UALBANY
State University of New York

Institute of Informatics
Logics and Security Studies

Evaluating Human-Machine Conversation for Appropriateness

Nick Webb
September 24, 2009

ILS Tech Report: 007

Available for download from:
`http://www.ils.albany.edu`

1 Abstract

As more sophisticated automatic conversational dialogue systems are created, the search for a reliable evaluation paradigm that goes beyond global dialogue scores, such as user satisfaction, or parameterisation over observable metrics, such as word error rate (WER) intensifies. Working within the EU funded COMPANIONS program, we investigate the use of appropriateness as a measure of conversation quality. We report initial work in this direction, including the annotation and comparison of the output of two generations of the same dialogue system.

2 Introduction

The development of COMPANION Embodied Conversational Agent (ECA) technologies requires new models of evaluation. With existing evaluation paradigms, it is often sufficient to track known dialogue system metrics (such as Word Error Rate (WER), Concept Error Rate (CER), levels of initiative and types and numbers of confirmation), and correlate these with global measures of user satisfaction captured using tools such as interviews, surveys and Likert scales.

Dialogue systems built and evaluated on these principles are certainly successful at their target tasks. However, there is a tendency when using these measures to learn dialogue behaviour to optimise over those parameters that introduce the most significant errors - often caused by Automatic Speech Recognition (ASR) - and offset this against ease of use. There is little account taken of the concepts of naturalness and appropriateness of dialogue - such as avoiding over confirmation or using information gained from the user in an appropriate way.

Further, Companions are targeted as persistent, collaborative, conversational partners. Rather than singular, focused tasks, as other deployed dialogue systems (cf. [Pec93, LRGB99, HBI$^+$05]), fully developed Companions can be expected to 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 the task itself is that of maintaining a relationship. When devising an evaluation paradigm for these systems, we need to balance the completion of tasks with some measure of "conversational performance", in the belief that managing and maintaining successful conversation will achieve our goals.

This conversational performance can be measured in terms of dialogue appropriateness - where this applies to both the functional aspects of the dialogue, and the content of the conversation. The functional aspect requires that the Dialogue Manager (DM) *does* the right thing at the right time, e.g. answers an immediate question, or identifies, and correctly processes clarification and error correction steps in dialogue. The content aspect requires that, whilst performing any given dialogue action, the DM *says* the right thing. This includes the notion of judging the phrasing of utterances, checking the referential content of phrases (using co-reference accurately, for example) and determining that questions asked are appropriate given current knowledge the DM can be assume to posses.

We propose to examine evaluations of dialogue appropriateness, to determine if they can be used to judge human-machine dialogue in this way, to judge if the information appropriateness annotates is useful, both to evaluators and developers, and to conduct evaluations of appropriateness over Companions prototypes, in an attempt to address three key goals:

1. To see if we can determine a trend occurring between appropriateness and user satisfaction (and hope that any such trend is positive)

2. To draw correlations between observable metrics, and the (time consuming and manually performed) appropriateness measures

3. To develop an evaluation paradigm for dialogue systems that, whilst remaining principally task oriented, has the additional goals of establishing and maintaining useful relationships.

For this paper, we report on initial appropriateness evaluation performed over the first Companions prototypes.

3 Related Work

There has been a significant amount of effort spent on evaluating spoken language dialogue systems (cf. [DG95, Min98, WRA$^+$02]), 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 are somewhat functional descriptors of quality of interaction.

The DARPA Communicator Program made extensive use of the PARADISE metric [WLKA97]. 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 - measures 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 (cf. [HM06]. 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.

4 Evaluating Companions

The motivation for developing new evaluation techniques was the on-going development of Companion technologies[1], the first prototypes of which include a personal automated 'Senior Companion' (SC), that learns about its owner's life story, needs, and preferences through free-ranging natural language dialogue with them, whilst offering the user appropriate kinds of help with IT tasks, suitable entertainment, and companionable conversation, 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. An interaction with the Senior Companion can be seen in Figure 1, and an interaction with the HNF can be seen in Figure 2.

In order to address the three key goals outlined in the introduction, we

---

[1] Being developed under European Commission Sixth Framework Programme, Information Society Technologies Integrated Project IST-34434 (`http://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

need to establish baseline measures, using existing and well-understood observable metrics, and initial user feedback. This first evaluation was conducted over the fall of 2008, and was principally concerned with initiating and testing the metric evaluation process. We evaluated three manifestations of the Companions concept the 'Senior Companion' (SC) [CDPW08], the 'Health and Fitness Companion' (HNF) [THS^{+}08] and the mobile HNF [SGH^{+}08] - a version of the HNF designed to work on mobile platforms and PDAs. The evaluation also served as a preliminary feedback mechanism for the development of the prototypes.

The mechanism for this evaluation was two-fold:

Metric-centric The use of quantitative methods to determine values for speech metric data including word error rate (WER) and concept error rate (CER) of the various prototypes. Alongside these are a variety of dialogue metrics such as dialogue duration; number of turns; words per turn, etc.

User-centric - Qualitative methods used to acquire subjective opinions from the users of the Companions prototypes, including Likert based surveys, focus groups and interviews.

The results of the Phase 1 evaluation are reported in [BHW08]. During this evaluation, all interactions with the Companions prototypes were collected, and are collectively available as an XML encoded corpus for use by evaluators and developers alike.

(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

5 Appropriateness

Several researchers [SF93, HT96] have looked at using a mechanism of appropriateness of dialogue as a measure of an effective communication strategy. We principally used the work of [TRS04] in evaluating the ICT/ISI mission rehearsal exercise system. This system is a training or tutorial system, where soldiers can run several training missions, and try a range of tactics to complete these missions. Although the domain is task-oriented, the main objective of interactions with the system is *not necessarily* efficient task performance, but to learn actions and consequences over the life cycle of the interaction. The user is the major source of knowledge with respect to the actions - so that even when the artificial agent knows what should be the next action, the user is given a wide degree of initiative, even if it is known, a-priori, that this leads to less efficient task performance. The key similarity here to Companion technologies is this balance between appropriate interactions and task completion. This appears to be a reasonable assumption for Companion-like conversation where the user is expected to have control over the dialogue in most situations. Of course, it is to be expected that the system take initiative when performing certain central tasks, but the control and flow of dialogue in these situations is relatively well understood. What remains the focus of our evaluation is allowing the user to control and direct the conversation as they feel appropriate, and have the system respond in an acceptable way.

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.

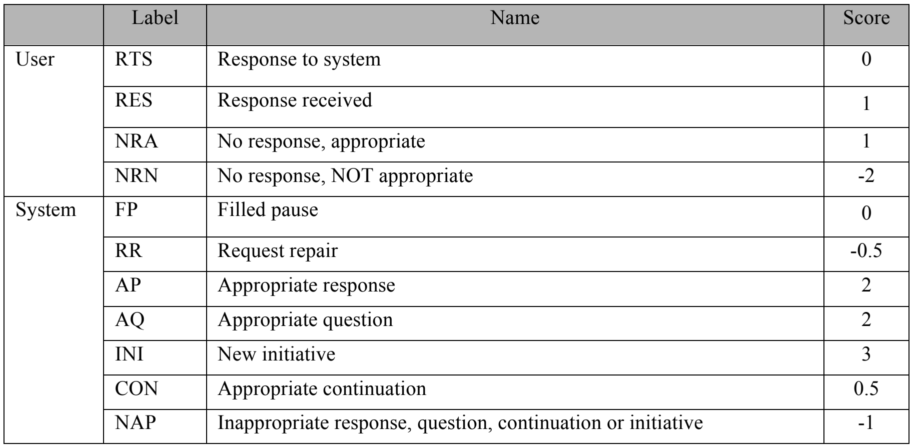
5.1 Mark-Up

In order to capture appropriateness of dialogue, annotation of the dialogue transcript is required. Annotators use a system (a version of which is outlined in [TRS04]) 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) and codes each with one of several annotations (a summary of which can be seen in Figure 3). These annotations are derived in part from the Gricean Maxims of conversation [Gri75], and several prior studies used some variation of a coarse application of these maxims where dialogue was marked as appropriate or not. [TRS04] added some finer distinctions which can be useful when assessing the behaviour of automatic systems.

For users, there are four grades of annotation - 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 trial 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 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). Examples of these categories as applied to the mission rehearsal system can be seen in [TRS04].

5.2 Scoring

Each of the resulting annotations over the transcript is then scored. Below 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



	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

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 modeling, it can be seen that such an evaluation methodology can be used to grade complete and part dialogues.

6 System Evaluations

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 known to the Campanion.
- 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 [FCC$^+$09].

From the metric evaluation, 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, for example, significantly more users finding the Companion 'engaging' *and* indicating that the Companion 'demonstrated emotion' at times.

6.1 Appropriateness Evaluation

With subjective and objective evaluations complete, we have XML files containing the all user-system interactions, and a measurable 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 5.1, and determine the score for each dialogue, and for each system as a whole, to see if we could characterise and capture improvements in the dialogue.

---

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

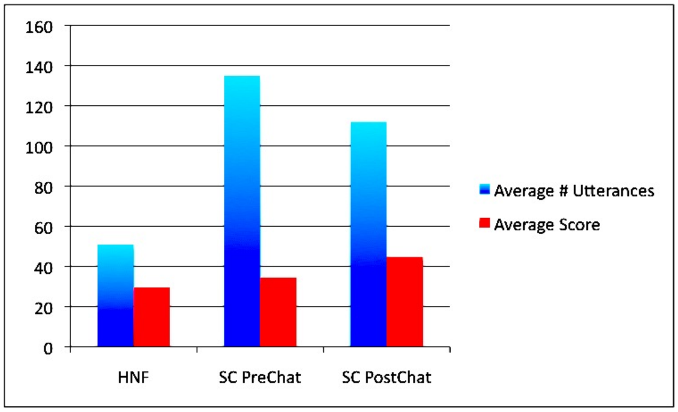


Figure 4: Average number of utterances, and average score, per system

For this evaluation, all dialogues were annotated by a single user. However, to check consistency, one dialogue of the HNF and one of the SC was annotated by 3 additional annotators, with no training other than the annotation guidelines[3]. 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 [Kri04]. 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

---

[3] available for download from `http://www.nick-webb.net`

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 8.

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.

6.2 Analysis

Once all dialogues had been annotated, we used the scoring mechanisms, outlined in Section 5.2, to calculate average scores for each system. It should be obvious that, with the current scoring mechanism, longer dialogues have greater chance of scoring higher than shorter dialogues, and indeed this appears to be the case when comparing dialogues between the HNF and SC. A summary of the average dialogue score, and a comparison with the average number of utterances can be seen in Figure 4. However, it is important to note that with the initial Companions prototypes, XML output files from the system are not *exactly* comparable - with important differences in the segmentation of system utterances. This means that the differential between HNF and SC average utterances is not as pronounced as it seems here. That said, there is some notion that, for Companion-like technologies, longer conversations are better (a metric often refered to as ‘stickiness’, and at odds with other task-oriented systems, where shorter interactions are better) - in-

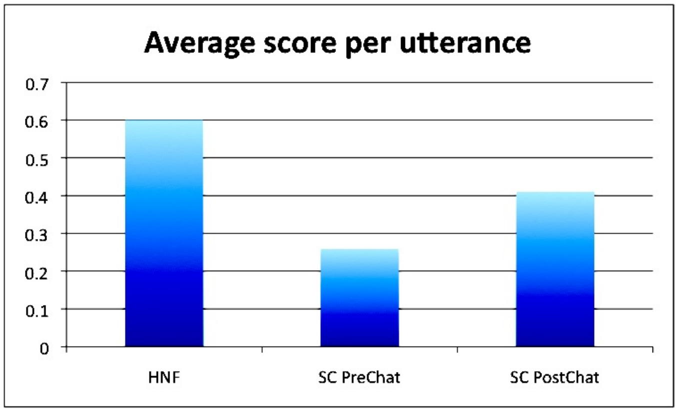


Figure 5: Average score of *each* utterance, per system

dicating 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 (talking about a potentially infinite number of photographs, versus a fairly well constrained domain of day planning) also has an effect.

If we try to 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 utterances of the HNF. Some combination of overall total score, and average score per utterance seems to be beneficial and we will explore that further in future work.

Where comparison is most useful at this 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. 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 “Edingburgh”, 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 cat-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.

7 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, in Section 6, 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.

8 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

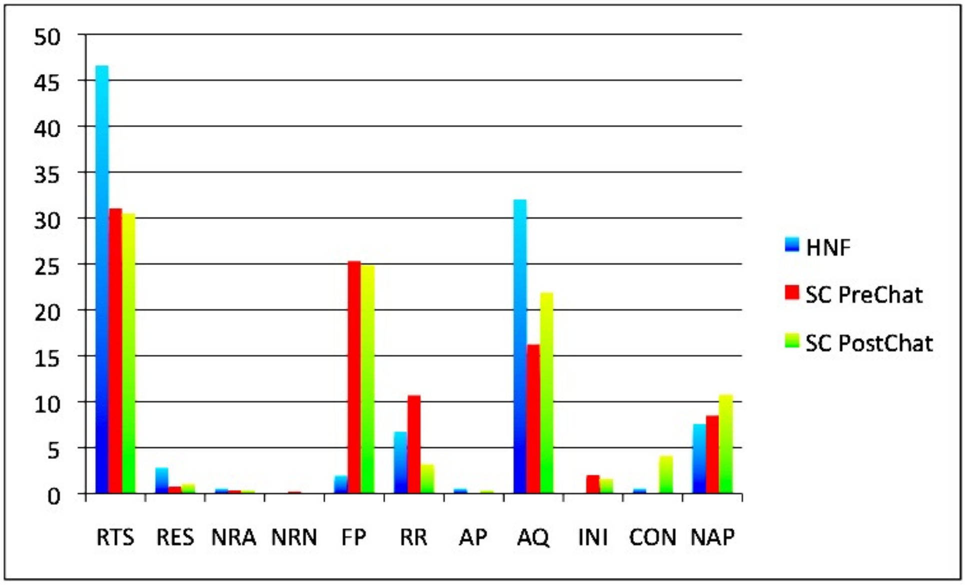


Figure 6: Distribution of tags (*as %*) across systems

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 [AGLT08]) that could work in combination 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.

References

[AGLT08] Ron Artstein, Sudeep Gandhe, Anton Leuski, and David Traum. 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, 2008.

[BHW08] David Benyon, Preben Hansen, and Nick Webb. Evaluating Human-Computer Conversation in Companions. In *Proceedings of the 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.

[CDPW08] R. Catizone, A. Dingli, H. Pinto, and Y. Wilks. Extraction tools and methods for Understanding Dialogue in a Companion. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008.

[DG95] Morena Danieli and Elisabetta Gerbino. 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, 1995.

[FCC$^+$09] Debora Field, Roberta Catizone, WeiWei Cheng, Alexiei Dingli, Simon Worgan, Lei Ye, and Yorick Wilks. The Senior Companion: a Semantic Web Dialogue System. In *Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems*, Budapest, Hungary, 2009.

[Gri75] Paul Grice. Logic and Conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics*, volume 3. Academic Press, 1975.

[HBI$^+$05] H. Hardy, A. Biermann, R. Bryce Inouye, A. McKenzie, T. Strzalkowski, C. Ursu, N. Webb, and M. Wu. The AMITIÉS System: Data-Driven Techniques for Automated Dialogue. *Speech Communication*, 48:354–373, 2005.

[HM06] Melita Hajdinjak and France Mihelič. The PARADISE Evaluation Framework: Issues and Findings. *Computational Linguistics: Special Issue on Empirical Studies in Discourse Interpretation and Generation*, 32:263–272, 2006.

[HT96] L. Hirschman and H. S. Thompson. 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, 1996.

[Kri04] Klaus Krippendorf. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Thousand Oaks, CA, 2nd edition, 2004.

[LRGB99] L. Lamel, S. Rosset, J.L. Gauvain, and S. Bennacef. The LIMSI ARISE system for train travel information. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[Min98] W. Minker. Evaluation methodologies for interactive speech systems. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, pages 199 – 206, Granada, Spain, 1998.

[Pec93] J. Peckham. A new generation of spoken dialogue systems: results and lessons from the SUNDIAL project. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pages 33 – 40, Berlin, Germany, 1993.

[SF93] A. Simpson and N. Fraser. Blackbox and glassbox evaluation of the SUNDIAL system. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 1993)*, Berlin, 1993.

[SGH+08] O. Ståhl, B. Gambäck, P. Hansen, M. Turunen, and J. Hakulinen. A Mobile Fitness Companion. In *Proceedings of the 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.

[THS+08] 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. Multimodal Agent Interfaces and System Architectures for Health and Fitness Companions. In *Proceedings of the 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.

[TRS04] David R. Traum, Susan Robinson, and Jens Stephan. 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, 2004.

[WLKA97] M. Walker, D. Litman, C. Kamm, and A. Abella. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Jan 1997.

[WRA+02] 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. DARPA Communicator evaluation: progress from 2000 to 2001. In *ICSLP*, 2002.