Evaluating Conversation with Hans Christian Andersen

Niels Ole Bernsen, Laila Dybkjær and Svend Kiilerich

Natural Interactive Systems Laboratory
University of Southern Denmark
Campusvej 55, 5230 Odense M, Denmark
nob@nis.sdu.dk, laila@nis.sdu.dk, kiil@nis.sdu.dk

Abstract

This paper presents an analysis of data from a large-scale in-field Wizard of Oz simulation of a conversational domain-oriented edutainment system. The basic turn-level data are described and the 10 longest conversations are analysed in order to evaluate the theory of domain-oriented conversation underlying the design specification of the system.

1. Introduction

Spoken language dialogue systems (SLDSs) which can pass the classical Turing test of computational intelligence [Turing 1950] are still a distant goal. We might call such SLDSs *domain-independent* systems, i.e. they would be capable of conducting spoken conversation about virtually any domain of discourse in the way humans do. Today's all-dominant SLDSs paradigm, on the contrary, is the *task-oriented* system. Such systems can be successfully developed and deployed because developers can use the powerful constraints provided by the task to apply a dialogue structure which ensures appropriate system responses to user input in, say, 95% of the input cases [Bernsen et al. 1998]. Interestingly, in-between task-oriented systems and domain-independent systems lies the halfway post of *domain-oriented* SLDSs. The basic challenge in developing domain-oriented systems is that the task constraints are gone because the system must be able to conduct spoken conversation about virtually any topic within its domain(s). On the other hand, the domain-oriented system is not required to handle conversation about topics outside its domain(s).

In the EU NICE project on Natural Interactive Communication for Edutainment, 2002-2005, http://www.nice-project.com), we are developing a domain-oriented SLDS which enables users to have spoken conversation with fairytale author Hans Christian Andersen (HCA). Over ten days in the summer of 2003 at the HCA Museum in Odense, we collected 30 hours of English spoken dialogue field data in a Wizard of Oz (WoZ) simulation of the first system prototype specification. This data, totalling some 500 conversations, has been analysed to evaluate the first prototype design. This paper presents the basic turn-level data and an analysis of the 10 longest conversations in order to evaluate the theory of domain-oriented conversation underlying the specification.

2. The NICE HCA system

In the NICE scenario, 3D life-like HCA is in his study surrounded by artefacts, furniture, and open floor space. A door leads into a fairytale games world populated by some of his fairytale characters, such as the Naked Emperor and Cloddy-Hans. Travelling the virtual world in first-person perspective, the user can have spoken conversation with HCA and use 2D gesture input for topicalising artefacts during conversation. At some point, the user is invited to visit the fairytale world to engage in spoken computer games with the characters. The system is designed for edutainment use in public spaces. The target users, children and adolescents 10-18 years old, are expected to use the system for 5-20 minutes each.

HCA's conversational ability is based on the *domains* he is familiar with: (1) his eventful life and experience; (2) his fairytales; (3) his person, perceived physical presence, study and its artefacts; (4) his "gatekeeper" role for access to the fairytale games world; (5) the user; and (6) the meta domain for handling meta-communication. HCA gathers knowledge about the user and uses the information during conversation. Each of HCA's domains are decomposed into relatively generic *topics.* In the Life domain, for instance, topics include greetings, HCA_identity, HCA_age, HCA_physical_appearance, HCA_study, and HCA_personality. Under, e.g., the HCA_study topic, HCA may talk to the user about objects in his study which the user indicates through gesture and/or speech. Domain-wise, the first prototype is broad and (relatively) shallow. All domains are included but HCA can only adequately address a fraction of each of them. The cover story is that HCA is back but still in training to become what he once was.

HCA has a rather simple emotional state-space model of being default friendly or more or less happy, angry or sad. He reacts emotionally to some user input topics and semantics, displaying his reactions verbally and through gesture, facial expression, etc. A finite state machine controls the HCA character module's three output states in which he is alone in his study working, expresses awareness of the user's current input, and outputs a conversational contribution, respectively [Bernsen 2003]. For a description of the implemented first prototype, see [Bernsen et al. 2004].

3. Wizard of Oz setup

The WoZ setup was the following. Two wizards took turns working in the basement of the HCA Museum. A student helper invited kids and youngsters to talk to HCA who was present on a laptop furnished with a headset. Figure 1 shows HCA as users saw him in the WoZ simulation. Input gesture was not available so it was not possible to travel the virtual world and point to objects. Also the fairytale games world and HCA's gatekeeper role were absent from the simulation. Using voice distortion, the wizards mostly followed the first prototype output specification which was navigable for output lookup on their screen. They were instructed to improvise conversation about some non-specified topics to gauge user interest

in topics outside the specification, i.e. their summer holidays, the museum, and explaining recent technical inventions to HCA. The wizards were also tasked with sometimes misunderstanding the input. All conversations were logged and transcribed.

Figure 1. HCA talking to a user.

4. Conversational principles

HCA's specified model of conversation is based on the following principles of prototypical successful human-human conversation: (1) search for common ground, i.e. shared knowledge, interests, etc. [Clark 1996], because (2) success depends on it; (3) interlocutor contribution symmetry in terms of activity and expertise-sharing; (4) expressive story-telling of, e.g., personal experiences, anecdotes, humour; and (5) the permissibility of rhapsodic topic-shifts on a baseline of coherence. Our claim is that (1) through (5) are essential to the successful design of domain-oriented conversation for edutainment.

Following presentation of the basic turn-level WoZ data in Section 5, Sections 6 through 8 propose how to evaluate the top-ten WoZ conversations as to their achievement of properties (1) through (5). Needless to say, as there is little experience available on how to evaluate a system such as the present one, we have had to develop new evaluation metrics.

5. Basic data

Table 1 shows the basic turn-level simulation data. Turn numbers show the total number of turns made by the user and HCA in a conversation. Since they take turns communicating, each of them will produce half of the turns +/- a single turn. No [age, gender] means that the users did not tell the system their age or gender.

The total of 498 conversations only excludes four conversations of <4 turns and two conversations in which the transcribers mixed up the users. The reason why Table 1 provides substantial information on users' age, gender and nationality, is that HCA has as a priority in conversation to gather this user information for conversational use. He will thus try to collect this information either up front or, at least, early on in each conversation. Roughly, age, gender, and nationality information was provided by 90% of the users. The most common reason, by far, for not providing age, gender, and/or nationality information was that the user broke off conversation before HCA could gather this data. Thus, the average turn numbers for no-age and no-gender users is as low as 13 and 14, respectively. In a few cases, the wizards forgot to ask for the information. Few users refused to tell HCA their age or gender, and only in a couple of cases is there reason to believe that a user gave deliberately wrong information. An example is Maria on Day 9 who first had a 98-turn conversation as Maria, an 11 years old female from Denmark, and then came back to have a 24-turn conversation as Maria, a 13 years old boy from Denmark wanting to discuss girls with HCA, unfortunately with limited success.

Table 1 shows a rather close gender balance of 210 (47.3%) female users and 234 (52.7%) male users, as well as near-identical turn averages for female and male users, i.e. 30 and 29, respectively.

Item	Totals	Item	Totals
Conversations	**498**	Av. turns <10	**26**
Age <10	**49**	Turns 10-18	**7563**
Age 10-18	**240**	Av. turns 10-18	**32**
Age >18	**164**	Turns >18	**4328**
No age	**45**	Av. turns >18	**26**
Male	**234**	Turns no age	**581**
Female	**210**	Av. turns no age	**13**
No gender	**54**	Turns male	**6689**
Countries	**29**	Av. turns male	**29**
Turns all	**13739**	Turns female	**6310**
Av. turns all	**28**	Av. turns female	**30**
Turns <10	**1267**	Turns no gender	**740**
Av. turns no gender			**14**

Table 1. Basic WoZ simulation data.

6. Common ground

Arguably, as shown in Table 2, the top-ten conversations demonstrate successful edutainment conversation in which the interlocutors have achieved substantial common ground (Section 4). Otherwise, the top-ten conversations would hardly have achieved a turn average of 109 turns. A more thorny issue is whether a target group turn average of 32 turns (38 turns for target group native English speakers) constitutes edutainment success. At this point, we do not know how to evaluate this figure.

7. Symmetry in conversation

According to our theory of conversation (Section 4), *symmetry* is a key property of prototypical successful human-human conversation. Symmetry (or balance) is manifested by, at least, (1) symmetry in presenting expertise in domains of common interest, (2) symmetry in taking initiative in changing the domain/topic of conversation, and (3) symmetry in being an active contributor in driving the conversation forward. We emphasise that symmetry, thus defined, is *not* a measure of the edutainment success of the system. In principle, depending on the age, personality, etc. of the user, the system may achieve edutainment success even with an extremely passive user (2,3) who lets HCA drive the conversation and is happy listening to stories from HCA's domains of expertise (1). Rather, we submit, *enabling*, or *developing for*, symmetrical conversation is a mandatory and difficult goal whose achievement will make the application likely to succeed with users having very different interests and personalities. A system which only, or primarily, caters for users who want to endlessly listen to HCA's stories, is a story-telling machine rather than a give-and-take conversational system.

7.1. Expertise and domain/topic change

Looking at the distribution of turns on user and HCA expertise in the top-ten conversations, we find the following. Showing in parentheses who is the expert, i.e. either the user (U), HCA (H), or both (B), the conversations addressed 16 domains/topics in total. Italics show domains/topics improvised by the wizards: Age HCA (H), games (U), greetings (B), *inventions* (U), H's knowledge (H), H's life (H), H's looks (B), *the museum* (U), out-of-domain (B), H's study (H), *travels* (H), user (U), *vacation* (U), *weather* (B), who is H (H), H's works (H). The number of turns spent on user and HCA areas of expertise was 485 and 520, respectively, yielding a balance of 0.93. This suggests that we have identified a reasonably symmetrical set of areas of conversation for the system. The most popular domains/topics, by far, were HCA's works (258 turns), *inventions* (189), games (134), the user (122), HCA's study (100), and HCA's life (83). Perhaps the main surprise is that HCA's life only comes in 6th place.

As regards the second symmetry metrics introduced above, i.e. who opens a new topic or phase of conversation among the 16 presented above, the top-ten conversations show a user/HCA initiative distribution of 59/76 or 0.78. This may not be quite enough for "full" symmetry, yet the symmetry measured in this way suggests that the combined conversational mechanisms used by the wizards are on the right track as regards users who find substantial common ground in conversation with HCA. It should be noted, however, that the balance measured includes a substantial number (20) of user-topic initiatives which are made to start or end the conversation. Even subtracting the, equally quasi-mandatory, HCA initiatives taken to collect user information (12), the 20 user greetings may be viewed as a distorting factor. Thus, the corrected symmetry metrics of 0.61 demonstrates that there is still important work to do to achieve symmetry in terms of initiative to open new topics of conversation.

7.2. Activity symmetry

Intuitively, the participants in conversation are equally active in driving the conversation forward if they show an equal distribution of (a) initiative in the conversation *and* (b) volunteered information. We thus propose to measure the extent to which a partner drives the conversation by the sum of that partner's initiative and volunteered information. The point is that measuring initiative-only is insufficient. Prototypical successful human-human conversation does not just consist in a mutual barrage of questions. It is also characterised by the participants feeling sufficiently at ease to volunteer information, personal and otherwise, for their partner(s) to comment upon.

Operationalising, we define two measures. *Initiative* is measured by: (i) the number of information requests made, including questions, requests to be told, and any other information directive. We term all information directives *questions;* (ii) the number of volunteered information items, or *observations*, offered. Thus, *activity symmetry* is measured as the proportion of user and HCA questions and observations.

Questions are pretty easy to identify in the corpus. In the metrics applied, we ignore user initiatives to start and end the conversation, including "hello", "how are you", "I must go now", and the like, since these conventional initiatives cannot be considered indicators of user initiative. We also exclude meta-communication directives, such as "what did you say?", requested repetitions, and anomalies such as unfinished questions whose point cannot be determined. However, *non-requested* repetitions, such as when the user insists on getting a question answered which HCA has somehow dodged, for instance by making an irrelevant observation, are counted as questions.

User *observations* must be *volunteered*, i.e., they cannot be direct replies to HCA's questions. Rather, they occur when HCA has *not* made a question but, e.g., a reply to a question, an observation, an interjection showing interest, such as "yes" or "interesting", has praised the user's knowledge or skills, or the like. A user observation can also form part of a reply turn, as in, e.g.: HCA: "Do you like the Little Mermaid fairytale?" User: "Yes. I have seen the cartoon movie". In this exchange, the remark about the cartoon movie is volunteered information. The user is not conversationally obliged to add anything in particular. Very often, the user simply says, e.g., "oh", "interesting", "yes" (or "no" if required in the context), or "cool". Such interjections are not counted as observations. To count as a (user) observation, the utterance must contain new and non-conventional turn-taking information, such as when a user, discussing cars with HCA, tells him that there is a car-racing facility in his neighbouring village. Moreover, if an observation is followed by a question in the same turn, which is something HCA often does but which users rarely do, we do not count the observation because it is unlikely to be reacted to.

id	QO	W	T	QO/T	B:U/H	W/T
1	2	161	49	**0.04**	**0.04**	**3.3**
HCA	47		49	**0.96**		
2	44	650	79	**0.56**	**0.93**	**8.2**
HCA	47		78	**0.60**		
3	19	541	93	**0.20**	**0.26**	**5.8**
HCA	70		92	**0.76**		
4	8	390	50	**0.16**	**0.19**	**7.8**
HCA	42		49	**0.86**		
5	14	356	53	**0.26**	**0.35**	**6.7**
HCA	39		52	**0.75**		
6	7	453	45	**0.16**	**0.18**	**10.1**
HCA	41		45	**0.91**		
7	5	271	41	**0.12**	**0.16**	**6.6**
HCA	30		41	**0.73**		
8	11	503	48	**0.23**	**0.35**	**10.5**
HCA	31		47	**0.66**		
9	5	388	40	**0.13**	**0.15**	**9.7**
HCA	34		39	**0.87**		
10	22	282	49	**0.45**	**0.59**	**5.8**
HCA	37		49	**0.76**		

Table 2. User activity.

Table 2 shows the results on user activity in the top-ten dialogues. Column 1 from the left shows the 10 users and HCA (H). Column 2: number of questions and observations per interlocutor. Column 3: words per user. Column 4: questions and observations per turn. Column 5: user/-HCA activity symmetry based on Column 4. Column 6: words per turn per user. The table makes it quite clear that, on average (0.32), drive symmetry is far more diffi-

cult to achieve that the symmetries discussed in Section 7.1. Moreover, individual variation is rather extreme, ranging from 6-year old Vaughan from the USA at 0.04 to Marius, Norway, at an almost perfect 0.93. This variation strongly suggests that edutainment value is not strongly linked to drive symmetry. Our detailed figures on each user also suggest that user age, gender, or English proficiency cannot be used to predict drive symmetry. It is probably a coincidence that the two native English speakers in the corpus, Vaughan, and Jenny, 17 years old, have the two lowest drive symmetry figures.

7.3. Symmetry of verbal activity

It is tempting to propose a fourth conversational symmetry metrics in terms of user/HCA words per turn averages. However, since HCA sometimes tells long stories implemented at design-time, this metrics makes little sense. All one has to do to change it is to make HCA tell longer stories. However, average user turn length *might* be considered a measure of user activity in conversation. Clearly, it does make a difference to our intuitive judgment of user activity if a user tends to make single-word contributions or speaks at length throughout. The average user turn length in the top-ten conversations is 7.5 words per turn. This average represents large variations among the users, ranging from Vaughan's 3.3 words per turn to Marie's 10.5 words per turn. The top-ten WoZ corpus suggests that there might be a gender difference in user verbal activity. In this small corpus, the average verbal activity for girls is 8.8 words per turn and for boys 6.1 words per turn.

Another interesting question is whether the top-ten conversations suggest any correlation between user drive symmetry and average turn length. Although Vaughan is lowest on both counts, generally, the conversations do not suggest any correlation. For instance, Marius and Maria, the top-two in drive symmetry, are not particularly verbally active. And Signe and Jenny, who are pretty low on drive symmetry, are close to the top in verbal activity.

8. Story-telling and topic shifts

HCA is designed to be able to tell stories about three of his fairy tales, his family, the pictures in his study, etc. He does so when conversationally relevant and the user accepts to hear the stories which is usually the case. The users' story-telling is mostly elicited by HCA asking the user for an explanation of some invention or game mentioned by the user. Other cases of user story-telling occur when HCA persuades a user to tell him about one of his fairytales or what the user does in his holidays. Only in rare cases does a user volunteer a story, such as when a user tells HCA that she was born in the countryside.

Topic shifts are a common phenomenon in conversation. In the conversations analysed, most topic shifts happen smoothly. When two persons meet for the first time, it is natural to exchange information about, e.g., name, age, and origin. Some users seem puzzled to be asked about their gender but accept the question, probably because they realise that even a life-like computer-animated character cannot see them. The transition between topics, such as different fairytales, HCA's study, games, and inventions, is usually smooth. When a topic is exhausted and none of the interlocutors have more to say about it, it is natural to move on to something different. Alternatively, a new topic may come up due to an association made while discussing some other topic. Thus, jumping to new topics is not a problem in itself. However, the challenging topic shifts are those which happen when the user addresses an out-of-domain (OOD) topic. HCA basically handles such situations in one of two ways in the analysed conversations. He either (i) utters a context-free HCA quote, e.g., "The emperor of China is Chinese" or "It was so lovely out in the country". The original idea behind (i) was to make the user change topic. In reality, such utterances rather make the users fall silent because they do not know what to say in return. Alternatively (ii), HCA asks a question about a different topic rather than replying. Some users try to re-iterate their unanswered question but typically give up after a couple of failed attempts.

9. Conclusion and Future Work

We have analysed data gathered in an in-field WoZ experiment in order to evaluate the conversation model of a domain-oriented system enabling conversation with fairytale author HCA. The analysis suggests that common ground is achievable with some fragment of the target users, at least, and that expertise symmetry has been reached. However, symmetry in domain/topic shifts and, in particular, activity, remain distant goals for which, moreover, we lack appropriate target figures. Finally, HCA's topic shifts in response to OOD input reveal a major challenge in improving conversational coherence.

Recently, a first prototype system with the WoZ-tested functionality (minus wizard improvisations) was tested with 18 users from the target user group. This gave us a first chance to ask users their opinion on the system. Although generally quite positive in many respects, the interviews demonstrate, as might be expected from the WoZ simulations, that conversational coherence remains a challenge. Our second prototype design will be guided by the WoZ and user test interviews, logfiles, and AV recordings.

Acknowledgement

The NICE HCA work is being supported by EU's Human Language Technologies programme under contract IST-2001-35293. We gratefully acknowledge the support.

References

Bernsen, N. O.: When H. C. Andersen is not talking back. In Rist, T., Aylet, R., Ballin, D., Rickel, J. (Eds.): *Proceedings of the Fourth International Working Conference on Intelligent Virtual Agents*, Irsee, Germany, 2003. Berlin: Springer Verlag 2003, 27-30.

Bernsen, N.O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D., and Mehta, M.: First prototype of conversational H.C. Andersen. To appear in *Proceedings of the International Working Conference on Advanced Visual Interfaces* (AVI 2004), Gallipoli, Italy, 2004.

Bernsen, N.O., Dybkjær, H., Dybkjær, L.: *Designing Interactive Speech Systems.* From First Ideas to User Testing. Springer Verlag, London, 1998.

Clark, H.H.: *Using Language.* Cambridge: Cambridge University Press, 1996.

Turing, A.: Computing machinery and intelligence. *Mind* 59, 1950, 433-60.