Iterative Development and Evaluation of a Social Conversational Agent

Annika Silvervarg, Arne Jönsson
Department of Computer and Information Science
Linköping University, Linköping, Sweden
annika.silvervarg@liu.se, arne.jonsson@liu.se

Abstract

We show that an agent with fairly good social conversational abilities can be built based on a limited number of topics and dialogue strategies if it is tailored to its intended users through a high degree of user involvement during an iterative development process. The technology used is pattern matching of question-answer pairs, coupled with strategies to handle: follow-up questions, utterances not understood, abusive utterances, repetitive utterances, and initiation of new topics.

Introduction

Social aspects of conversations with agents, such as small talk and narrative storytelling, can have a positive effect on peoples general interest in interacting with it and help build rapport (Bickmore, 2003). It can also be utilised to develop a relationship and establishing trust or the expertise of the agent (Bickmore and Cassell, 1999). We are interested in exploring if and how these and other effects transfer to an educational setting where children and teenagers interact with pedagogical agents in virtual learning environments. We see several reasons to incorporate social conversation with such agents, for example, it allows for cognitive rest, it can increase overall engagement and receptivity and it can make students feel more at ease with a learning task or topic (Silvervarg et al., 2010). There has, however, been few attempts to understand the users' behaviour in social conversations with pedagogical agents (Veletsianos and Russell, 2013) and embodied conversational agents (Robinson et al., 2008).

In this paper we report on how we iteratively have worked with addressing the questions of 1) what do users talk about during social conversation with a pedagogical agent, 2) how do users talk during social conversation with a pedagogical agent, 3) how does the answers to 1) and 2) affect the dialogue functions needed to implement social conversation with a pedagogical agent.

A social conversational pedagogical agent

Our work extends a virtual learning environment with an educational math game named "The Squares Family" (Pareto et al., 2009). A crucial part of the environment is a pedagogical agent, or more specifically a teachable agent (Biswas et al., 2001). While the student is playing the game, the agent "learns" the rules of the game in two ways, by observation or through on-task multiple choice questions answered by the user. A teachable agent is independent and can act on its own, yet is dependent on the student to learn rules and strategies. The intended users are 12-14-year-old students, and the teachable agent is designed as having the same age or slightly younger.

The conversational module for off-task conversations has been developed as a rather independent module of the learning environment. Off-task conversation is based on a character description of the teachable agent that is consistent with the overall role of the agent as a peer in the environment.

The challenge can be seen as a question of managing the students' expectations on the agent's abilities. Our approach was to frame and guide the interaction with the student in such a way that, ideally, the shortcomings and knowledge gaps of the agent never become a critical issue for achieving a satisfying communication. We have therefore chosen to work with user-centred agile system development methods to be able to capture the users' behaviour and tailor the agent's conversational capabilities to meet their expectations. This includes combining focus group interviews and Wizard-of-Oz role-play (Dahlbäck et al., 1993) with development and evaluation of prototypes, surveys and analyses of natural language interaction logs.

The off-task conversation is implemented using a slightly extended version of AIML, Artificial Intelligence Markup Language (Wallace, 2010). AIML works on the surface level and map user utterances to system responses. User utterances can consist of words, which in turn consist of letters, numerals, and the wildcards _ and *, which function like words. Synonyms are handled using substitutions and grammatical variants through several different patterns for the same type of question and topic.

Responses consist in their simplest form of only plain text. It is also possible to set or get data in variables and predicates, give conditional responses, choose a random response from a set of responses, and combinations of these. AIML also allows for handling a limited context by either referring to the systems last utterance or a topic that span multiple exchanges.

Prototype 1

In the first iteration an agent persona was developed through focus groups with 20 target users. The persona sketch formed the basis for WOz-style role play, in which students simulated off-task conversations in the game. Three students played the part of the agent, and four students played the role of the user. The resulting 12 dialogues were analysed according to topics, linguistic style and dialogue phenomenon. A number of new topics emerged that had not been brought up in the focus groups. The linguistic style of the utterances could be characterised as grammatical, short sentences, with the use of smileys and "chat-expressions". The dialogue mostly consisted of unconnected question and answer pairs, but some instances of connected dialogue with 3-4 turns occurred. The initiative was evenly distributed between user and system. There were frequent use of elliptical expressions, mostly questions of the type "what about you", but no anaphora.

Based on these findings the first prototype implemented basic question-answer pairs, a strategy for follow-up questions from the agent and user, a topic model with 6 topics that could be initiated by the agent (to allow for mixed-initiative dialogue), and a very simple strategy to handle utterances that the agent could not understand. To handle variations in user input (choice of words and grammatical variations) the system used substitutions where, for example, synonyms and hyponyms were substituted for a "normalised" word, and variations of patterns that used the "normalised" keywords. The agent's replies were sometimes randomly chosen from a set of 3-5 variants to get some variation if the user asked the same question several times. Follow-up questions where randomly attached to half of the answers the agent gave to questions from the user. When the agent did not understand a user utterance it said so three out of four times, but in one out of four it instead initiated a new topic and posed a question to the user. The agent also initiated a new topic when the user gave an acknowledgement, such as ok, after an answer from the agent.

To evaluate the system a total of 27 students tested the prototype. After a short introduction to the project and the system they played the game for 10 minutes, chatted with the agent for 5 minutes, then played the game for 5 minutes and chatted for 5 minutes again. Analysis of the corpus showed that failed interpretations had to be dealt with. Many of the failed interpretations were due to linguistic variations on known topics, and most of all acknowledgments, but also greetings and follow-up questions. Topics also needed to be expanded, both new topics, for example age, food, pets, favourite colour and cars, but also more sub-topics related to, for example, computer games, school subjects, and TV. Topics initiated by the agent were proved to be a good strategy and implied that the model should be expanded with new topics and more sub-topics. If the agent could initiate more topics it would both make the dialogue more fun but also help the user to know what type of topics the agent could talk about. A recovery strategy to deal with utterances that the system still were not able to interpret correctly, for example utterances out of scope, was also needed to help the user understand what they could talk about and avoid losing track of the conversation. The recovery strategy should also include ways of breaking long sequences of utterances that included, for example, random letters and "Ok" sub-dialogues, and steer the dialogue back to a topic.

Evaluation of the prototype also showed that hazing, testing and flaming occurred rather frequently, for example comments and questions regarding the agent's appearance, intellectual capacities, and sexual orientation and activities. To fend this off, the agent also needed to recognise such utterances and have a strategy to direct the dialogue

in a more productive direction.

Prototype 2

The second prototype expanded the agent's dialogue functionality with a recovery strategy for utterances not understood by the agent, a strategy to handle abusive utterances, a strategy for delayed follow-up questions by the user (e.g. Agent: How are u? User: Fine. Agent: Ok. User: wbu?), an extended topic model, a better model for agent initiative and flow of the dialogue, and extended vocabulary.

The recovery strategy for utterances not understood by the agent was implemented similar to the one used by SGT Star (Artstein et al., 2009); first a simple "clarification", utterance 2 in Figure 1, then, a "stall", followed by a "move on", and finally a "prompt" that directs the user to a new topic. In our system the agent stall is a general prompt for the user to initiate a new topic, utterance 4 in Figure 1, if this fails the agent proposes a topic, utterance 6 in Figure 1, and as a last resort takes the initiative and asks a question, utterance 8 in Figure 1.

1	User:	fgd
2	Agent:	I did not understand that
3	User:	cb
4	Agent:	I did not catch that, can we talk about something else?
5	User:	s
6	Agent:	Let's talk about school
7	User:	s
8	Agent:	My school is okay, a little too much homework, wbu?

Figure 1: Example of how the agent handles utterances that are not understood.

The strategy to handle abusive utterances are similar in that the first response is general, such as "eh, let's talk about something else", followed by two attempts to change the topic, randomly done either by a general request for the user to suggest a topic or by the agent to introduce a new topic, followed by a remark that further abuse will result in a report to the teacher. If the user continued with abusive utterances the loop starts again

To avoid repetitive utterance sequences such as many greetings, laughters or acknowledgements in a row, the agent initiated new topics when those types of utterances where repeated. The AIML parameter topic was used to handle delayed follow-up questions from the user. The topic model used for this purpose was extended to include a total of 15 topics, where some were related, for example music and favourite artist.

Evaluation of the prototype was conducted in the same way as for prototype 1. This time with 42 users, 22 girls and 20 boys. Analysis of the chat logs revealed that the model for follow-up questions needed to be revised. Since follow-up questions were initiated randomly by the agent it sometimes asked for information the user already had provided, which seemed to irritate the user. The model for topic initiation by the agent could also be refined to provide more coherence in the dialogue. Another problem detected was that the strategy to use the current topic as context to interpret generic follow-up questions sometimes was overused and led to misunderstandings when the user tried to introduce a new topic. The agent thus needed a more sophisticated strategy to handle topics and topic shifts.

Prototype 3

The main improvements of prototype 3 was the introduction of mini narratives, an improved strategy for follow-up questions and an improved strategy to introduce and continue a topic. For three main topics, free time, music and school, sub-topics where added, and when the agent took the initiative it tried to stay within topic and either tell a mini-narrative or introduce a sub-topic. Follow-up questions where now only posed if the user had not already provided answers to the question earlier in the conversation.

The conversational agent was evaluated at a Swedish School, where 19 students, from three classes, 12-14 years old, used the learning environment with the conversational agent during three lectures. The students played the game for about a total of 120 minutes and after every second game session a break was offered. During the first three breaks the students had to chat with the agent until the break ended, after that chatting was optional.

Table 1 shows the proportion of different types of user utterances in the logged conversations. The coding scheme is based on the coding schemes used by Robinson et al. (2010) to evaluate virtual humans. As can be seen in Table 1 most user utterances are "appropriate" in that they are either Information requests (Q), Answers (A), General dialogue functions (D) or Statements (S), but a total of 22% are "inappropriate", i.e. Incomprehensible

(G) or Abusive (H).

Table 1: Dialogue action codes and proportion of different agent utterances.

Code	Description	Prop
D	General dialogue functions, e.g. Greeting, Closing, Politeness	14%
H	Hazing, Testing, Flaming, e.g. Abusive comments and questions	11%
Q	Information Request, e.g. Questions to the agent	31%
R	Requests, e.g. Comments or questions that express that the user wants help or clarification	0%
A	Answer to agent utterances	18%
S	Statements	16%
G	Incomprehensible, e.g. Random key strokes or empty utterances	11%

As for the agent's responses it seems that the system handles most utterances appropriately although many of these are examples of requests for repair, see Table 2. The highest value 3, i.e. appropriate response, means that the agent understood the user and responded correctly. Request Repair, is when the system does not understand and asks for a clarification or request that the user changes topic. Partially appropriate, code 2, is typically used when the user's utterance is not understood by the agent, and the agent's response is to initiate a new topic. Inappropriate response, code 1, is when the system responds erroneously, typically because it has misinterpreted the user's utterance.

Table 2: Agent response codes and proportion of different agent responses.

Code	Description	Prop
3	Appropriate response	51%
2	Partially appropriate	15%
RR	Request Repair	30%
1	Inappropriate response	4%

Given a definition of Correct Response as any response that is not inappropriate, code 1 in Table 2, we see that prototype 3 handles 96% of the user's utterances appropriately or partly appropriate. The proportion of responses where the system correctly interprets the user's utterance is, however, only 54%, and there are still 11% Flaming/Hazing which also affects the number of repetitions, which is very high. Most of the not correctly interpreted utterances and the repetitions, occurs when the student is hazing/flaming or testing the system, e.g. none of the user's utterances in Figure 1 is correctly interpreted (code 3) but all are correctly responded to (code 2).

Prototype 4

The evaluation of prototype 3 did not indicate any need for more sophisticated dialogue functions but rather that the number of correctly interpreted utterances needed to increase. Therefore the focus of prototype 4 was to add and refine patterns used for interpretation of user utterances, for example adding more synonyms and expressions. It also included adding answers to some questions related to the already present topics, for example, questions on the agent's last name and questions and comments about game play. Since prototype 3 still had problems with a lot of abusive comments prototype 4 also included a revised strategy to handle abusive utterances, where the agent gradually tries to change the topic and finally stops responding if the abuse continues to long. If and when the user changes topic the strategy is reset.

The evaluation of prototype 4 comprise conversations with 44 students, 12-14 years old. The students used the system more than once which gives us 149 conversations with a total of 4007 utterances of which 2003 are from the agent. Each utterance was tagged with information about, dialogue function, topic, initiative, agent interpretation, agent appropriate response, and abuse.

Many of the utterances that the agent cold not correctly interpret in prototype 3 were due to the fact that users did not engage in the conversation and did not cooperate, rather they were testing the agent, abusing it or just writing nonsense. We believe that the strategies we have developed to handle such utterances are more or less as good as a human. For the evaluation of prototype 4 we therefore modified the criteria for tagging an utterance as appropriate. An utterance was only appropriate if the agent responded as good as a human, taking into account that if a user utterance is very strange, a human cannot provide a very good answer either, see Table 3. In this new coding scheme we also removed the previous category RR where utterance that request repairs falls into R3 (if a human could not interpret the user utterance neither) or R2 depending on how appropriate they are in the context.

There are also cases when the agent's response may have been better if it was a human, but where it is not obvious how, or even that a human could

Table 3: Agent response values.

Code	Value
R3	Agent responses that a human could not have done better
R2	Agent responses that are ok but a human may have responded better
R1	Agent responses that are erroneous because the agent did not understand the student or misunderstood

do better. We tag these R2 as well not to give credit to the agent for such responses.

Table 4 shows topics with information on how many utterances in total belonged to each topic, and how well the agent responded to utterances within each topic (R1, R2 or R3), as well as the proportion of not appropriate or only partially appropriate responses in percentage. NO TOPIC is for utterances like greetings, requests for repair, random letters or words, and abuse. As can be seen in Table 4 the agent gives appropriate responses (R3) to 1399, i.e. 70%, of the users' utterances. Table 4 lists all the topics present in the corpus and shows that although given the opportunity to talk about anything, users tend to stick to a small number of topics.

Table 4: Topics present in the corpus and the number of appropriate responses (R3), partially appropriate response (R2), and non-appropriate responses (R1).

TOPIC	**Tot**	**R3**	**R2**	**R1**	**Prop R1 + R2**
NO TOPIC	534	527	50	6	10%
PERSINFO	317	189	147	32	56%
MUSIC	267	199	43	25	25%
SCHOOL	201	136	48	17	32%
FREE-TIME	177	136	35	6	23%
MATH-GAME	122	43	74	5	65%
COMP-GAME	103	80	19	4	22%
FOOD	38	15	18	5	61%
FAMILY	34	15	17	2	56%
FRIENDS	30	8	20	2	73%
MOVIES	24	19	3	2	21%
SPORT	21	12	6	3	43%
MATH	20	13	7	0	35%
ALCOHOL	5	2	3	0	60%
BOOKS	2	0	1	1	100%
CLOTHES	2	0	2	0	100%
FACE-BOOK	2	2	0	0	0%
PET	2	2	0	0	0%
TV	2	1	1	0	50%
Total	**2003**	**1399**	**494**	**110**	**30%**

To further investigate the utterances causing problems we looked at the responses tagged as R1 and R2 and classified them as caused by greetings (GREETING), questions (QUESTION), statements (STATEMENT) or utterances where correct interpretation depends on the dialogue history (HISTORY). The proportions of problematic utterances and the dialogue functions of these utterances are shown in Table 5.

Over half of the problematic utterances are questions. Of these the majority are regular questions, while 30% of them are specific follow up questions on a previously introduced topic. A small number are generic follow up questions either directly following an answer to a question posed by the agent (Agent: Do you like school?, User: yes, wbu?), or a free standing delayed question (Agent: Do you like school? User: Yes. Agent: ok, User: wbu?). Statements are causing 29% of the not appropriate answers, mainly statements and answers to questions. There are also some abusive comments and random utterances. Problems related to the dialogue history is comparatively small. It includes both answers, statements and different kinds of questions. Examples of utterances the agent cannot handle well are follow up questions on topics previously introduced by the agent or the user, statements that comment on previous answers, use of anaphora referring to previous questions or answers, users' attempt to repair when the agent does not understand, and delayed answers to questions asked more than one utterance before.

From Table 6 we see that most of the problems relate to a small number of topics. PERSINFO, FREETIME and MATHGAME have mainly problems with statements and questions. The agent has for example insufficient knowledge and ability to talk about the math game itself. It also lacks knowledge about personal information such as hair colour, eye colour and other personal attributes. MUSIC and SCHOOL are common topics where the user often tries to make follow up topics that the agent cannot handle.

Conclusions

We have worked iteratively with user centred methods and rather straightforward natural language processing techniques to develop a social conversational module for a pedagogical agent aimed at students aged 12-14 years. The importance of involving students in the development process cannot be underestimated. Initially they

Table 5: Type of utterances that causes not appropriate responses, and their dialogue function.

	R1	R2	Tot	Prop
GREETING	**2**	**15**	**17**	**2,8%**
Greetings	2	15	17	2,8%
QUESTION	**72**	**244**	**316**	**52,6%**
Questions	45	141	186	30,9%
Specific Follow up Questions	21	74	95	15,8%
Generic Follow up Questions	1	17	18	3,0%
Answer + GFQ	1	8	9	1,5%
Abuse	2	5	7	1,2%
STATEMENT	**17**	**158**	**175**	**29,1%**
Statement	9	71	80	13,3%
Answer	4	45	49	8,2%
Acknowledgement	2	16	18	3,0%
Abuse	1	16	17	2,8%
Random	1	8	9	1,5%
HISTORY	**19**	**74**	**93**	**15,5%**
Answer	3	25	28	4,7%
Statement	2	26	28	4,7%
SFQ	4	10	14	2,3%
Random	8	3	11	1,8%
GFQ	1	4	5	0,8%
Question	1	4	5	0,8%
Acknowledgment		2	2	0,3%

Table 6: The distribution of different types (G: Greetings, H: History, Q: Questions, S: Statements) of problematic utterance for different topics.

TOPIC	G	H	S	Q	Tot
PERSINFO		11	38	130	179
MATHGAME		9	23	47	79
MUSIC		17	30	21	68
SCHOOL		23	21	21	65
NO TOPIC	19	10	21	6	56
FREETIME		6	13	22	41
COMPGAME		2	8	13	23
FOOD		1	7	15	23
FRIENDS		1	1	20	22
FAMILY		1	7	11	19
SPORT			5	4	9
MATH		2	1	4	7
MOVIES			2	3	5
ALCOHOL			1	2	3
CLOTHES				2	2
BOOKS				2	2
TV				1	1
Total	**19**	**83**	**178**	**324**	**604**

gave us valuable insights on the capabilities of an agent capable of social conversation. In the iterations to follow they provided feedback on how to refine the conversation to handle both ”normal” conversation as well as not so conventional conversation. Using questionnaires to measure system performance or as an instrument for further development is not fruitful (Silvervarg and Jönsson, 2011). We have instead relied on analysis of the logs to find bugs, and detect patterns that suggest lack or sophistication of dialogue capabilities that should be added or refined.

The strategy has been fairly successful. We seems to have captured what users talk about very well. The number of topics is surprisingly small given that the user can introduce any topic they want. A possible improvement could be to include a more elaborate model for topics and subtopics for some topics. There are also still knowledge gaps concerning some questions within topics, such as personal attributes and traits of the agent.

How they talk about the topics are also fairly well understood, in that the dialogue capabilities needed have been discovered and implemented. It may be that addition of anaphora resolution could improve the agents responses, but that would probably be a marginal improvement, since problems related to anaphora are very rare. Some of the problems are related to the large variation of how the same question or statement can be expressed, and the limited power of interpretation based on keywords, but this does not seem to be a big problem. The same can be said for spelling mistakes. Inclusion of an automatic spellchecker may increase the successful interpretations, but probably only to a small degree.

A remaining problem that is hard to address is the fact that some users are very uncooperative. They deliberately test the system or are just not engaging in the dialogue but rather write nonsense or abuse. Previous studies have shown that there seem to be three types of users (Silvervarg and Jönsson, 2011): 1) those that really try to use the system and often also like it, 2) users that do not use the system as intended, but instead tries to find its borders, or are bored and never tries to achieve an interesting dialogue, but rather resorts to flaming/testing/hazing, and 3) those that are in between. Users of type 1 are rather unproblematic, as long as the agent has enough topics and sub-topics they will have a meaningful conversation. Users of type 2 will probably never be engaged in a meaningful conversation with the agent no matter how sophisticated it is. Focus must instead be to avoid users of type 3 to adhere to type 2 behaviour, which could be achieved by having a variety of techniques to handle abusive and testing behaviour and enough topics and sub-topics to allow for a varied enough conversation, as presented in this paper.

References

Ron Artstein, Sudeep Gandhe, Jillian Gerten, Anton Leuski, and David Traum. 2009. Semi-formal evaluation of conversational characters. *Languages: From Formal to Natural*, pages 22–35.

T Bickmore and J. Cassell. 1999. Small talk and conversational storytelling in embodied interface agents. In *Proceedings of the AAAI Fall Symposium on Narrative Intelligence*.

T. Bickmore. 2003. *Relational Agents: Effecting Change through Human-Computer Relationships*. Ph.D. thesis, Media Arts & Sciences, Massachusetts Institute of Technology.

G. Biswas, T. Katzlberger, J. Brandford, Schwartz D., and TAG-V. 2001. Extending intelligent learning environments with teachable agents to enhance learning. In J.D. Moore, C.L. Redfield, and W.L. Johnson, editors, *Artificial Intelligence in Education*, pages 389–397. Amsterdam: IOS Press.

Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of oz studies – why and how. *Knowledge-Based Systems*, 6(4):258–266. Also in: *Readings in Intelligent User Interfaces*, Mark Maybury & Wolfgang Wahlster (eds), Morgan Kaufmann, 1998.

Lena Pareto, Daniel L. Schwartz, and Lars Svensson. 2009. Learning by guiding a teachable agent to play an educational game. In *Proceedings of AIED*, pages 662–664.

Susan Robinson, David Traum, Midhun Ittycheriah, and Joe Henderer. 2008. What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In *Proceedings of LREC 2008*, Jan.

Susan Robinson, Antonio Roque, and David R. Traum. 2010. Dialogues in context: An objective user-oriented evaluation approach for virtual human dialogue. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Annika Silvervarg and Arne Jönsson. 2011. Subjective and objective evaluation of conversational agents in learning environments for young teenagers. In *Proceedings of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

Annika Silvervarg, Agneta Gulz, and Björn Sjödén. 2010. Design for off-task interaction – rethinking pedagogy in technology enhanced learning. In *Proceedings ot the 10th IEEE Int. Conf. on Advanced Learning Technologies, Tunisia*.

George Veletsianos and Gregory Russell. 2013. What do learners and pedagogical agents discuss when given oppportunities for open-ended dialogue. *Journal of Educational Computing Research*, 48(3):381–401.

Richard S. Wallace. 2010. Artificial intelligence markup language. URL:http://www.alicebot.org/documentation/.