

Iterative Development and Evaluation of a Social Conversational Agent

Abstract

We present results from the iterative development and evaluation of a social conversational agent. The conversation takes place in a virtual learning environment. In this paper we focus on analyzing what about and how users talk during social conversations, and the strategies necessary to implement a social conversational agent. The conclusion from our work is 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 the development process.

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 (for an extended discussion see Silvervarg et al. (2010)).

As stated by Robinson et al. (2008) there has been few attempts to understand the users' behaviour in social conversations with embodied conversational agents and:

With the more ambitious goal of creating agents with more truly conversational abilities, a very central measure of progress will necessarily be based on both discovering and utilizing sufficient information about the central interests and language of the target audience.

In this paper we report on how we iteratively have worked with addressing the questions of i) what do users talk about during social conversation with a pedagogical agent, ii) how do users talk during social conversation with a pedagogical agent, iii) how does the answer to i) and ii) 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). The game specifically trains basic arithmetic skills, such as carry-overs and borrowings, with focus on grounding base-10 concepts in spatial representations. 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.

A 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 a teachable agent that is consistent with the overall role of an agent as a peer in the environment. The

starting point for the design of the social conversation with the agent were the following design challenges:

1. how to deal with students' expectations regarding the agent's knowledge and social competencies,
2. how to deal with students' varying degree of engagement in social interaction with the agent, and
3. how to deal with the risk of verbal abuse known to arise when students interact with conversational agents (Veletsianos et al., 2008).

The challenges 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 with development and evaluation of prototypes, surveys and analyses of natural language interaction logs. For the present project, the intended users are 12-14-year-old students, and nearly 200 students in this age group have taken part or are still taking active part in the development. A short description of the iterations and evolving prototypes follows.

Iterative development and evaluation

The development of the social conversational module have been done in three main iterations. Each iteration has led to addition of more topics and extended vocabulary, but also more sophisticated strategies for dialogue management.

Prototype 1

In the first iteration an agent profile was developed through focus groups with 20 target users. A persona sketch formed the basis for a 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 persona had the same content but the agents

had different gender, names and pictures. Agent players were asked to act in accordance with the persona, and in case dialogue topics occurred outside its known scope (for example about pets), they were asked to improvise. 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, for example, literature and where the agent lived. 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.

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. This is used to correctly respond to follow-up questions and answers to questions posed by the agent.

The agent's replies are often randomly chosen from a set of 3-5 variants. To deal with recurring types of utterances, such as greetings, hazings, and flammings a number of variables are used to keep track of repetitions. To be able to choose new topics the agent has a topic model implemented as a set of AIML predicates including topics that are linked to questions or narratives.

The resulting agent was capable of mixed initiative dialogue, could handle generic follow-up questions by both user and agent (e.g. you, wbu),

acknowledgments, (e.g. ok, aa, mm, I see), and had a simple strategy to choose topics for agent initiatives. There was a very rudimentary model with 6 topics linked to questions and the possibility to store students' answers related to these topics. The student could either introduce the topics and the agent could ask a follow up question, or the agent could initiate a new topic, for example, when it failed to interpret a student's utterance. See Figure 1 for a dialogue that illustrates the capabilities of the agent.

- 1 User: How ar you
- 2 Agent: What do you do in your freetime
- 3 User: Play computer games
- 4 Agent: ok
- 5 User: wbu
- 6 Agent: hang out with friends
- 7 User: oh
- 8 Agent: ok
- 9 User: okay
- 10 Agent: ok

Figure 1: Example of a dialogue between a user and the first prototype.

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 min, chatted with the agent for 5 min, then played the game for 5 min and chatted for 5 min 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 introduced 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 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.

Prototype 2

The second prototype expanded the agent's dialogue functionality with a better model for agent initiative, extended vocabulary, an extended topic model, a strategy for error handling, a strategy to handle abusive utterances, and a strategy for delayed follow-up questions (e.g. Agent: How are u? User: Fine. Agent: Ok. User: wbu?). These strategies are described and exemplified below.

Prototype 2 were designed to allow for a mixed-initiative dialogue where both agent and user could take the initiative and initiate new topics and ask general follow up questions, e.g. "what about you". The agent took the initiative in situations where interaction was repeated or the user did not take the initiative, for example, when greetings, acknowledgments or laughters are repeated, see utterance 4 in Figure 2. Taking the initiative could result in both questions or continuation of the current topic through mini narratives, as exemplified in utterance 10 in Figure 2.

- 1 User: Hi
- 2 Agent: Hi
- 3 User: Hello
- 4 Agent: What type of music do you like?
- 5 User: much u?
- 6 Agent: everything but dance band music hehe
- 7 User: hehe
- 8 Agent: lol
- 9 User: lol
- 10 Agent: I listen to rock and rnb I think that's the sweetest music

Figure 2: Example of the agent taking the initiative.

A recovery strategy to deal with utterances that the system is not able to interpret correctly was implemented in prototype 2 to help the user understand what they can talk about and avoid losing track of the conversation. The strategy used is similar to the one used by SGT Star (Artstein et al., 2009); first a simple "clarification", utterance 2 in Figure 3, 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 3, if this fails the agent proposes a topic, utterance 6 in Figure 3, and as a last resort takes the initiative and asks a question or tells a narrative, utterance 8 in Figure 3.

Evaluation of prototype 1 showed that hazing, testing and flaming occurs rather frequently, for

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

example comments and questions regarding the agent's appearance, intellectual capacities, and sexual orientation and activities. To fend this off, the agent recognised such utterances and used a strategy similar to the repair strategy where the first response is general, such as utterance 2 in Figure 4, 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, utterances 4 and 6 in Figure 4, followed by a remark that further abuse will result in a report to the teacher, utterance 6 in Figure 4. If the user continued with abusive utterances the loop starts again.

1 User: you are ugly
 2 Agent: eh, let's talk about something more interesting
 3 User: you smell
 4 Agent: Let's talk about music
 5 User: ugly
 6 Agent: come on, let's talk about something fun
 7 User: ugly
 8 Agent: Stop that, or I will tell a teacher

Figure 4: Example of how the agent handles abuse.

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%

The evaluation of prototype 2 used a coding

scheme focussing on the agent's conversational abilities. The coding scheme is based on the coding schemes used by Robinson et al. (2010) to evaluate virtual humans. It has a set of codes characterising the user's dialogue action and another set of codes that evaluates the agent's responses. We only used a subset of the codes in the top layer used by Robinson et al. (2010) since our focus was on the quality of the agent's answers and we thus had no need to further differentiate the different utterances made by the users. Table 1 shows the categories and descriptions used for action coding. Each user utterance is coded with one of the categories in Table 1.

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%

The agent's responses are categorised as 3, 2, *RR*, or 1, Table 2, depending on the appropriateness of the response. It is a modified version of the coding scheme used by Robinson et al. (2010). Since we do not have any instances of "No response" we end up with four instead of five values for the agent's response. Figure 5 illustrates some of the codings. 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, see the dialogue in Figure 5. Inappropriate response, code 1, is when the system responds erroneously, typically because it has misinterpreted the user's utterance.

For coding of the agent's utterances we also use a category for agent initiatives, *I*, and one for repeated agent utterances, *REP*. The category *I* is used *only* when the system deliberately takes control of the interaction from the user, for example, posing a question on a new topic after a repeated sequence of user utterances that the agent is unable to interpret, see Figure 5.

The conversational agent was evaluated at a Swedish School, where 19 students, from three

Speaker, Utterance	Code
User are you free on Tuesday	Q-RR
Agent I did not understand what you said, can we talk about something else	
User hi	D-3
Agent Hi	
User are you good at school	Q-3
Agent rather good	
User neerd	H-3
Agent maybe we can talk about hobbies instead	
User I like football	S-RR
Agent what, how do you mean?	
User I like kicking the balls	S-RR
Agent hmm, I have difficulties understanding what you say, can we talk about something else	
User do you have a problem with Swedish	Q-2
Agent this weekend I am going to the movies. It is great to forget everything and just disappear into another world, at least if it is a good film, how about you?	
User may I join	Q-2
Agent What are you doing at the weekends?	

Figure 5: Example dialogue with taggings from the coding scheme.

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. 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).

As for the agent's responses it seems that the system handles most utterances appropriately, see Table 2, although many of these are examples of RR, the agent very seldom (4%) responds inappropriately, code 1. Given a definition of Correct Response as any response that is not inappropriate, code 1 in Table 2, we see that prototype 2 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 repe-

titions, occurs when the student is hazing/flaming or testing the system, e.g. none of the user's utterances in Figure 3 is correctly interpreted (code 3) but all are correctly responded to (code 2).

Prototype 3

The evaluation of prototype 2 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 3 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 2 still had problems with a lot of abusive comments prototype 3 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. Unfortunately this does not always work. For an example where the agent can not handle abuse see Figure 6.

- 1 User: you are ugly
- 2 Agent: eh, let's talk about something fun
- 3 User: you smell
- 4 Agent: come on, let's talk about something else
- 5 User: ugly
- 6 Agent: maybe we can talk about school instead
- 7 User: ugly
- 8 Agent: knock it off, or I don't want talk to you anymore
- 9 User: ugly
- 10 Agent:
- 11 User: ugly
- 12 Agent:

Figure 6: Example of how the agent non-successfully handles abuse.

The evaluation of prototype 3 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. The dialogue functions were refined to include more fine-grained categories than before, including specific follow up question (SFQ), generic follow up question (GFQ), and acknowledgement (Ack). The topics were based on known topics implemented in the agent, but the coding scheme was

also expanded with new topics that occurred in the corpus.

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

Many of the utterances that the agent could not correctly interpret in prototype 2 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 3 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.

- 1 User: your tactics has improved
- 2 Agent: I rather talk about something else

Figure 7: Example of erroneous handling, R2, of STATEMENT on topic MATHGAME

The tagging scheme is stricter than the one used when evaluating prototype 2, see Table 2. Using that scheme an utterance such as utterance 2 in Figure 7 would have been considered appropriate given that the agent could not fully understand utterance 1 (the conversation before has mainly been a greeting session). A human, however, is believed to understand that this is a statement praising the agents math game abilities and ought to acknowledge that with a *Thank you* or something similar.

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 do better. We tag these R2 as well not to give credit to the agent for such responses. Figure 8

is a typical such example. Utterance 2 in Figure 8 is just a random keystroke by the user and has no meaning at all. Maybe a human could capture this and inform the user that the utterance did not make sense, but it is not obvious how to do that.

- 1 Agent: What are you doing on weekends?
- 2 User: br
- 3 Agent: hmm, let us talk about something else

Figure 8: Example of utterance tagged R2 with no obvious improvement strategy

A typical example of an agent response tagged R1 is seen in Figure 9. Again the agent did not understand, but here the response is not acceptable.

- 1 User: do you like women
- 2 Agent: I use to play Age of Empires, what about you?

Figure 9: Example of erroneous handling, R1, of QUESTION on topic PERSINFO

Table 4 shows topics with information on how many utterance 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.

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. Typical examples of QUESTION and STATEMENT mistakes are illustrated in Figures 7 and 9. Figure 10 depicts an example of an error that is caused by the agent’s limited ability to remember the dialogue history. Utterance 6 could have been better responded to by a human as a human would know that utterance 5 refers to ”Smalltown” but our agent have not recorded that.

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%

- 1 User: where do you live
- 2 Agent: In Smalltown
- 3 User: bugger what a lousy town
- 4 Agent: what do you mean?
- 5 User: that it is bad
- 6 Agent: I do not understand, ask something else

Figure 10: Example of erroneous handling, R2, depending on lack of HISTORY

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.

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%

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 12-15 year old. The importance of involving students in the development process cannot be underestimated. Initially they 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"

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

conversation as well as not so conventional conversation. Using questionnaires to measure system performance or as an instrument for further development is not fruitful (Silvervag 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 by some 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 (Silvervag and Jönsson, 2011):

1. Those that really try to use the system and often also like it. They never, or seldom, use abusive utterances such as hazing or flaming and seldom test the systems capabilities. Instead they mostly try to be cooperative.
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.
3. Those that are in between. Often they start out "seriously" but after a while, maybe as a result of the system's performance, become more and more frustrated and enters in the second groups' behaviour.

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, however, 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.

In our studies the social chat have been a mandatory part of the use of the learning environment. Perhaps a study where the user could decide if they wanted to have a social conversation and how long it should last would give very different and probably "better" results, in terms of proportion of appropriate responses. For the future we plan to incorporate the social conversation in a learning environment where the user can choose to use it more freely as an alternative among many different activities. We are also planning on introducing some pedagogical interventions in the social conversation, for example, feedback on learning tasks, and questions that prompts reflection and metacognition about learning.

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.
- 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 Sjäöden. 2010. Design for off-task interaction – rethinking pedagogy in technology enhanced learning. In *Proceedings of the 10th IEEE Int. Conf. on Advanced Learning Technologies, Tunisia*.
- George Veletsianos, Cassandra Schrabar, and Aaron Doering. 2008. When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with computers*, 20(3):292–301.
- Richard S. Wallace. 2010. Artificial intelligence markup language. URL:<http://www.alicebot.org/documentation/>.