Smart Learning Environments

RESEARCH Open Access

Intelligent intervention by conversational agent through chatlog analysis

Mike Procter[1*], Fuhua Lin[1] and Bob Heller[2]

* Correspondence: mprocter@athabascau.ca
[1]School of Computing and Information Systems, Athabasca University, Athabasca, Canada
Full list of author information is available at the end of the article

Abstract

E-learning systems based on a conversational agent provide the basis of an intuitive, responsive, engaging interface for the online learner. This paper proposes an approach to intelligent intervention and strategic pedagogical design for improving student engagement when chatting with a conversational agent. First, we used previous conversational logs to detect and classify interaction behaviors of learners. And then we designed a set of strategies for intelligent intervention to improve learners' engagement when conversing with the conversational agents.

We implemented a multiagent framework to apply the strategy-based intervention. The effectiveness of learner interaction behaviors and the impact of intelligent intervention by the conversational agent were evaluated through chatlog analysis. Although not all of the quantitative tests were sensitive enough to detect the effect of the interventions, the findings suggest that the detection of behaviours was accurate. The interventions were observed to have the desired effect on behaviours associated with conversational engagement.

Keywords: Conversational agent, Online learning, Behavior analysis, Engagement intelligent intervention

Introduction

Conversational agents (CAs) are designed to provide learners with the ability to interact with computer software using natural language. In effect, the learner can chat with an application to obtain information or carry out tasks, receive coaching, practice a language, learn a new skill, or simply converse for social interaction or companionship. CAs embedded within e-learning applications have the potential to provide an intuitive, learner-friendly interface that engages the learner. Educational applications of CA technology include animated pedagogical agents (Heller and Procter 2009; Johnson et al. 2000), intelligent tutoring systems (D'Mello et al. 2008), collaborative learning (Kumar and Rosé 2011), and game-based learning systems (Bellotti et al. 2011; McClure et al. 2013).

The potential for education-related applications is particularly significant to e-learning as they can often be made remotely accessible via the Internet (Danforth et al. 2009; Heller and Procter 2011), or deployed on home computers and mobile device (Perez-Marin and Pascual-Nieto 2011) providing learners with on demand access to one-on-one and collaborative e-learning resources, available 24 × 7. Learners can also interact with embodied CA's participating in 3D virtual worlds (Heller 2016). An interesting role for CAs is to act the part of interviewee (Castellano et al. 2009). For example, for medical learners to practice their diagnostic skills against a virtual patient (Danforth et al. 2009).

In our previous research (Heller and Procter 2011), we developed a conversational agent for online learning psychology, Freudbot, to simulate a historical figure, Sigmund Freud. The underlying pedagogical design is based on the idea of using a series of narratives that can be delivered as a conversation. Learners converse with Freudbot, using text input, as if in the role of interviewing him. Freudbot is designed to respond in first person to questions and comments about Freud's life, family, theories, and colleagues. In all over 90 topics, broken into multiple narrative chunks, are programmed to be delivered to the learner following basic rules of conversation, such as greeting, turn-taking, and repairing misunderstandings. Thus, the learner can control the narrative in a natural way through the types of questions or comments they make while conversing with the CA. We believe that this involves a higher level of cognitive processing (Graesser et al. 2002; Yamashita et al. 2005). For this study, discourse features associated with engagement were developed as an extension of previous work in social presence (Heller 2016). Engagement has been linked to social presence (Castellano et al. 2009; Wise et al. 2004).

The purpose of this research is to investigate ways to improve the interaction between learners and CA for e-learning through maintaining learner engagement. We first propose and develop a methodology for learner engagement detection by analyzing the conversational log with text-analysis techniques to evaluate the quality and nature of the conversation by the learner. Based on the identified conversational behaviors of the learners, intervention plans are derived and applied as conversational strategies for maintaining, increasing, or re-establishing learner engagement. To evaluate the effectiveness of the proposed methodology, we developed a new Freudbot and conducted an experiment using students in an online psychology course.

The remainder of the paper is organized as follows. Section 2 provides a review of the literature on engagement-aware conversational agents. Section 3 introduces the methods for classifying learner conversational behavior, behavior detection, and intervention strategies. Section 4 details the experimental settings and a statistical analysis of conversational log data, which are discussed in Section 5. Finally, Section 6 summarizes the conclusions of the research and describes directions for future research.

Literature review

There are many issues associated with achieving this goal of improving the engagement of CAs. Several factors that can influence the perception of having a human-CA interaction that resembles that of a human-human one. Different approaches have been proposed to achieve these results. CAs can be embodied to give human-like characteristics and provide non-verbal cues through facial expressions and animated gestures (Cassell 2001; Johnson et al. 2000). Speech recognition and generation provides a human mode of communication. Becker et al. (2007)discuss modeling and expressing emotion and personality, while Kapoor and Picard (2005) focus on detecting learner emotion, and the ability to react appropriately to various affect states. Also important is the ability to recognize personality traits, and how they may affect the interaction (Mairesse et al. 2007).

Engagement-aware applications are able to detect the user's level of engagement and use this information to modify its actions or to change how the user interacts with the software. Detection or estimation of levels of learner engagement is fundamental to building engagement-aware application interfaces (Nakano and Ishii 2010; Xu et al. 2013). Techniques for detecting and identifying learner engagement typically require some method of sensing various verbal and non-verbal behaviour cues (Szafir and Mutlu 2012) as well as contextual information (Castellano et al. 2009). Eye-tracking headsets (Szafir and Mutlu 2012) are a common method for evaluating where the learner's attention is focused. Electroencephalography (EEG) headsets that can measure brain activity associated with engagement are becoming more prevalent, though researchers continue to search for ways to collect this data in a non-intrusive way (Asteriadis et al. 2009), a requirement for making engagement-aware applications a practical reality. Stereo cameras can detect body positioning and movement as well as expressive gestures (Xu et al. 2013).

Perhaps one of the least intrusive approaches to evaluating engagement is the analysis of conversation in dialogue based systems. A search of the literature found few references to this concept, despite the substantial amount of research associated with text-based affect detection. Wen et al. (2014) describe a technique for measuring cognitive engagement -- the degree to which one interprets or reflects upon a concept as opposed to simple description -- based on Turney's level of word abstraction dictionary (Turney et al. 2011) to distinguish between forum posts which are more descriptive and those that are more interpretive. The 2015 version of the Linguistic Inquiry and Word Count (LIWC) tool detects the number of words classified as 'analytical', which appears to be a similar measure as Wen's cognitive engagement, according to research reported in (Pennebaker et al. 2014). Our research attempts to break some new ground by proposing methods for detecting learner engagement based on their contribution to the conversation. This is done by using both a real-time analysis and an offline examination of the conversational record of learners.

As expected, the way in which learner engagement data is used is dependent on the application domain, the goals of the system, and role that learner engagement plays in the success of the interaction. Engagement information can drive strategies for persuading learners by advice giving applications (Novielli 2010), select appropriate strategies for reengaging inattentive learners (Szafir and Mutlu 2012), or simply make an embedded agent be perceived as more human-like (Castellano et al. 2009; Xu et al. 2013). Veletsianos and Russell (2013) feel that social discourse plays are important to the effectiveness of pedagogical agents. "The focus on task-oriented agents in the literature is in contrast to the vision of participatory, learner-centered, and community-oriented learning experiences" (Veletsianos and Russell 2013). For the purpose of this study we define engagement in terms of "conversational engagement" - our own term - which is the level to which the student is acting as a participant in the act of conversing with the CA. We provide examples of different levels of participation. This is the basis for applying the interventions.

The proposed approach

To enable a CA for e-Learning to adapt to information sources about the learner and to recover learner engagement, we first explore how to evaluate whether the learner is engaged in the conversational when interacting with a CA. An easily accessed information source is the conversational record, referred to as the "chat log", of the learner's interaction with the CA.

Classifying learner conversational Behaviours

Classifying learner input and CA responses

Two approaches to evaluating the conversation were investigated. The first involved creating a machine learning classifier that was trained on past conversational logs with the ITS, annotated with ratings for conversational quality and appropriateness of the user responses. The conversational logs from a previous study (Heller and Procter 2009) provided data from 10 min conversations by 90 participants chatting with the historical Figure CA, Freudbot. For Procter, Lin, & Heller (2016), we developed a *coding scheme* to classify the *learner input* and the *CA responses* in the conversational logs, and identified the following two key features for the *learner input* that are associated with level of learner engagement:

(1) *Response Appropriateness*: answering questions, responding to requests, addressing the topic under discussion, or changing to another domain related topic.

(2) *Conversational quality*: playing the role of conversant: using full sentences or phrases, not lone keywords, gibberish or random characters, and non-repetitive utterances.

Each student response was manually coded for the quality of conversation with values from 1 to 3. Response 1 represents what one would expect during a conversation, while 3 would be considered strange and inconsistent in a conversations. Response 2 was assigned if the coder was unsure. Each input was also rated for appropriateness on a scale from 1 to 4 based on how the student response compared to the ITS response. While initial performance figures are encouraging, this approach is still in the process of being developed. In particular, it was determined that more training data is required for the examples of poorer ratings.

Three patterns of learner behaviors

For the second approach we manually generated algorithms to categorize student input by identifying conversational behavioral patterns. This method has been shown to be reasonably robust. Identifying problematic conversational behaviors allows for a targeted form of intervention which attempts to repair or improve the conversation. This paper will focus on the second approach. We examined the logs of past studies using Freudbot (Heller, Procter, & Rose, 2016; Heller and Procter 2009) and identified three recurring patterns of learner behaviour:

> ***Tryer***: The learner attempts to ask questions exactly as one would hope they would, using full sentences (or close) on topics related to Freud. They continue to do this despite little or no success in getting Freud-related information from the CA. This *trying* behaviour is characterized by relatively long sentences, high

number of no-match cases per inputs and possibly input words with high abstractness value, a measure of cognitive engagement (Wen et al. 2014).
Keyworder: The learner answers questions or responds to bot output with single words or phrases associated with Freud or psychoanalysis, e. g. "ego", "psychoanalysis", "anxiety". Typically jumping from one topic to the next. This *keywording* behaviour could be detected by short inputs, non-repetition, low number of no-match cases per inputs, and possibly low abstractness value of input words.
Morer: The learner discovers a word that leads to advancement through the narrative and repeats that word. For example, just keeps saying "ok". *Moreing* behaviour could be detected by recognizing backchannel type words and phrases ("more", "ok", "I see"), and frequent consecutive repetition of those words.

Behavior detection

Learners may exhibit more than one of these behaviours. They may start off *trying* and eventually give up and start *moreing*. Or they might just stick with one strategy, like *keywording* and never experience a proper conversation. Often, these behaviours come about because of poor performance on the part of the CA, and the learner attempting to find a strategy that results in useful information being returned. Again, special functions have been programmed to identify certain learner dialogue acts, such as *backchannel* comments, which are used in conversation to indicate that one is following along and encouraging the other conversational partner to continue (e.g. "Okay", "I see", "uh huh"). Freudbot is programmed to recognize these phrases and continue the narrative associated with the current topic. The agent keeps a history of the use of these words and determines if consecutive repeated use of the same term has been used. In a similar way, Freudbot checks whether the learner is a *tryer*, indicated by longer sentences, suggesting complex questions or comments, followed by repair statements from the CA indicating it does not understand the learner input. The poor performance of the CA is an important aspect because an intervention is not required if the CA is successfully responding to the learner input with appropriate educational content. Again, if occurrences of this situation exceed a threshold, the associated data is published by a data source agent and received by a model agent. Another algorithm is used to detect potential *keyworder* behaviour. In each case, if the behaviour is detected enough times to exceed predetermined thresholds, the appropriate learner label – *tryer*, *morer*, or *keyworder* – is applied and this determination is published to the information stream, for the learner model agent to collect, possibly integrate with other data, such as the conversation quality, and determine if it should be passed on to the agent responsible for initiating interventions.

To tune and select the best parameters for the conversational behaviour detection algorithms we manually rated 26 conversations (613 turn pairs) from the chat logs of a previous experiment (Heller and Procter 2009) Each conversation was assigned a rating for each of the three types of behaviour: *trying*, *keywording*, and *moreing*. False positives were judged to have a negative effect since they are likely to trigger inappropriate interventions. This can be confusing to the learner, and undermine the perception of intelligence that plays a large part in engaging the learner. Results from comparing the manual and automated ratings were used to find the best balance between catching the behaviour and not accidently triggering a false intervention.

To measure the accuracy of the algorithms, chat logs from the current study (see Section 4) were manually coded to identify the three behaviors. The human coder read the entire log for a participant and assigned any behaviors observed, and a confidence rating from 1 (low) to 3 (high) for each behavior.

The agent's behavior assignments were compared against those of the human coder. Observations with low confidence ratings were ignored. As anticipated, the algorithms minimized false positives at the expense of false negatives, resulting in relatively high values for precision and relatively low values for recall (Table 1). Accuracy ratings are included but because there was a significant class imbalance for each of the behaviors it is potentially misleading as a performance measure. (Of 56 participants, manual coding found 48 tryers, 8 keyworders, and 21 morers.) F-scores, the harmonic mean of precision and recall, provide an indication of whether the balance of the two is reasonable. The F0.5 score is considered more appropriate because it weights recall lower than precision (by attenuating the influence of false negatives) which is consistent with the design objective of avoiding false positives ahead of reducing false negatives.

Interventions

We don't measure level of engagement, but instead detect behaviours associated with poor (e.g. *keywording*, *moreing*), and good (e.g. *trying*) conversational engagement. In Procter, Lin & Heller (2016) we describe how the behaviour detection algorithms were implemented as software agents which parse and analyze the conversation in real-time to evaluate the learner's conversational behaviour. The detection of any of the three conversational behaviours triggers an appropriate conversational intervention. The CA Representation (CA-REP) agent is responsible for monitoring events from the detection agents, and can direct the CA to inject an intervention into the conversation. We refer to this agent as the Intervention agent in this paper. The interventions support the pedagogical design described in Section 1 by encouraging the learner to make full use of the interactive narrative and conversational interface when it is determined that the student is not conversing or not exploring the narrative. The three behaviours and associated interventions are described briefly in Table 2.

The problem, simply put, is that the learner is either not managing to get to the CA content, as in the case of the *tryer*, or is not doing so through a conversational approach (*morer* and *keyworder*). The first type of problem is serious, the second is not optimal because the learner doesn't make use of the conversational capabilities of the CA. Although *morer* behaviour does expose significant Freud content, it is not much different than reading a book. *Keywording* is like using a search engine. Both cases leave little motivation for the learner to interact again. Both cases would likely result in a poor rating of the CA.

Table 1 Algorithm performance

	Accuracy	Precision	Recall	F_1	$F_{0.5}$
Tryer	0.702	0.919	0.708	0.800	0.867
Keyworder	0.912	0.714	0.625	0.667	0.694
Morer	0.807	1.000	0.421	0.593	0.784

Table 2 Behaviour types and interventions

Behaviour	Description	Intervention Number / Description	
Tryer	Attempts to use proper conversation but CA does not match most input	1	CA apologizes and suggests topics based on learner's area of interest (Freud's life or theories)
Keyworder	Does not attempt to converse. Enters single words or short phrases	2	Suggests conversational phrases to advance further into topics ("Tell me more about…")
Morer	Advances through topics by repeating the same "more" type word ("ok", "more", "go on")	3	Reminds learner they can branch to other topics ("Tell me about") or come back to a topic ("Tell me more about…")

If the learner can obtain content through a conversational approach, then there is no need to change anything. The learner is left in control of the conversation. If the system can recognize that the learner is having trouble obtaining content through a conversational approach, i.e. a *tryer*, the Intervention agent can address this by directing the CA to take some control of the conversation to introduce relevant topics. While this assumes some control from the learner, it is preferable to the learner having to resort to other behaviors to obtain useful information, such as just saying 'yes' (*moreing*), or using non-conversational input such as keywords.

In the case of *trying* behaviour, rather than stating "I don't understand" (or similar 'default' response), the CA 'recognizes' the problem and asks a question: "I don't seem to doing very well in trying to understand your comments and questions. If I can ask, are you more interested in my theories, or in my life?". The CA uses the learner's response to suggest an appropriate topic (theories, life/people, or both depending on stated preference). Additionally, future "no-match" responses will favor repair strategies that suggest topics related to the learner's interest, or ask leading questions related to the learner's interest. These are repair strategies that take away some of the control of the conversation from the learner, but are more likely to result in information being delivered.

In the case of *keywording* behaviour, the normal response from the CA will have the intervention appended to it ("I can't help noticing you have a somewhat abrupt conversational style. In any case, you can ask me to tell you more about a topic if you'd like to go into more depth."). The intention is to at least encourage the learner to use conversational directives to experience the narrative structure and appreciate the depth of the content, rather than just seeing the first section of each topic.

In the case of *moreing* behaviour, the process of triggering an intervention is the same as for *keywording*, i.e. the intervention is appended to a normal response. It informs the learner "You seem to be advancing the conversation by repeating the same word. This does allow you to cover a topic thoroughly, but remember that you can branch off to other topics ('Tell me about...') and come back to a topic ('Tell me more about…')." Again, the intention is to provide the learner with other ways to interact and encourage them to do so in a conversational way.

A secondary potential benefit of the interventions is to suggest that the CA has some level of awareness (of the learner's behaviour) and therefore promote a sense of social presence.

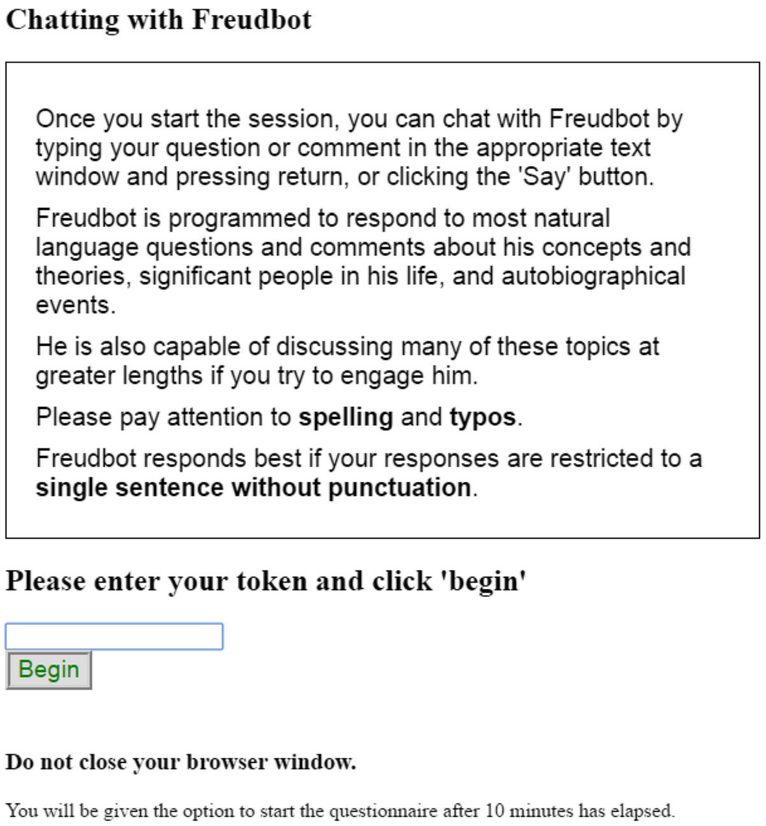
Experimental design and analysis

For system evaluation, the main purpose of the study was to assess the two decision support mechanisms provided by the agent-based system, which modify the behaviour of the CA so it could respond appropriately to detected learner conversational behaviours. The first expected outcome was a measure of the actual and perceived effectiveness of the interventions which are triggered by the agent system to be carried out by the CA. A second important outcome was to measure the effect of no-match conditions, where the CA is unable to classify the learner input to produce a meaningful response, and the perceived appropriateness of the strategy recommended by the agent system for the CA to employ. The other important goal was to gain some insight into what factors are related to learners perceiving the system as useful for learning, and which factors may drive the motivation to use it again. The two sources of data for analysis are the conversational record, or chat logs, and the responses to survey questions.

To evaluate the proposed approach, a total of 56 participants were recruited from a pool of undergraduate-level learners enrolled in an introductory psychology course. Students in this course are given the option of participating in a research study for course credit, for providing an experiential learning task associated with research methods. Participants were required to chat with Freudbot for at least 10 min. No direction was provided in terms of what to talk about, though some basic instructions were provided to optimize the interaction, as shown in Fig. 1.

Chatlog analysis

To answer the question whether or not the interventions provided by the system invoke a change in learner conversational behaviour, or experience, the chat logs are examined before and after each intervention, testing for differences in key variables. In



Chatting with Freudbot

Once you start the session, you can chat with Freudbot by typing your question or comment in the appropriate text window and pressing return, or clicking the 'Say' button.

Freudbot is programmed to respond to most natural language questions and comments about his concepts and theories, significant people in his life, and autobiographical events.

He is also capable of discussing many of these topics at greater lengths if you try to engage him.

Please pay attention to **spelling** and **typos**.

Freudbot responds best if your responses are restricted to a **single sentence without punctuation**.

Please enter your token and click 'begin'

Do not close your browser window.

You will be given the option to start the questionnaire after 10 minutes has elapsed.

Fig. 1 Freudbot start page

all, 37 participants received an intervention for *trying*, 7 for *keywording*, 8 for *moreing*. Thirteen (13) received no intervention. Note that some participants had more than one intervention, and some did not receive any.

The conversational logs contain timestamped records of exchanges, where an exchange is defined as input from the learner and a response from the CA. In addition, there is participant identification codes and special codes added within the conversation to show communication between the CA and the Intervention agent, including instructions to insert an intervention.

A custom application was used to pre-process the logs and output some summary data which we refer to as the Conversational Feature dataset. This includes the number of exchanges before and after intervention type, the number of cases where the CA could not interpret the learner input (no-match) for each intervention type, the number of cases where Freud-related content was output before/after each intervention, as well as ratios of all of these values per number of exchanges.

Intervention 1 – 'Trying'

For Intervention 1, the expected outcome is an improvement in the pedagogical utility of the experience, i.e. the delivery of more educational (Freud related) content. There is no attempt to change the learner's behavior because they are already conversationally engaged. The issue is that, generally due to the CA's shortcomings, the learner is not being rewarded for their efforts with useful information. The strategy is to, as naturally as possible, take some control of the conversation and provide content, while maintaining a conversational approach. Ideally this approach should be modeled after the way a human would handle the same situation. Faced with questions from an interviewer, but not understanding what they are getting at, one solution is to attempt to narrow down the area of interest, and then suggest a topic.

A paired-samples T-Test was carried out to compare Freud content and no-match counts before and after each intervention. The expected outcome is an increase in content delivered and possibly a decrease in number of no-match conditions. Measures of this are provided by the conversational features dataset and the results (Table 3) show that there is a significant increase ($p = .000$) in content and a significant decrease ($p = .009$) in non-matches after the delivery of Intervention 1. The effect size for Freud content (Cohen's d = .801) suggests a large effect (Cohen, 1988). The effect size for no-matches (d = .453) is small to medium. There were no significant differences in number Freud content or no-match exchanges before and after Interventions 2 or 3.

Table 3 Freud content and no-match counts before and after Intervention 1

Paired Samples Test

			Paired Differences				
	Mean	Std Dev	Mean	Std Dev	t	df	Sig (2-tailed)
Content before	.3789	.1422	−.1485	.1855	−4.867	36	.000
Content after	.5274	.1372					
No-match before	.2720	.1778	.0770	.1698	2.758	36	.009
No-match after	.1950	.1126					

Count = Number of Content/No-Match exchanges (before and after) per total exchanges (before ad after)

For interventions 2 and 3 The conversational logs were analyzed, in part, using the lexical analysis tool, LIWC. LIWC parses text, identifying and counting different components (e.g. number of words over 6 letters), and different parts of speech (nouns, verbs, articles, pronouns, etc), as well as linguistic dimensions, such as interrogatives, comparisons, and quantifiers. An important feature of LIWC is its use of proprietary dictionaries to classify words into different categories associated with psychological constructs, including affective, social, and cognition processes. The 2015 version of LIWC provides some new variables, including "analytical thinking" (Pennebaker et al. 2015).

The parsing application also formatted the log text for processing by LIWC. The participants' input was extracted, removing all other information, including Freudbot responses and timestamps. Special delimiters were inserted in the the text to identify where interventions had taken place.

Table 4 lists the LIWC output variables that were selected for analysis of the chat logs. These variables are intended to provide a measure of social presence and are based on those used in Heller and Procter (2014). The variable names changed slightly with the 2015 version of LIWC used in this paper.

Intervention 2 – 'Keywording'

For Intervention 2, the desired outcome is to affect a change in behavior in the way the learner is interacting with the CA, to boost conversational engagement, or at least provide the learner with the feedback to allow them to try a more conversational approach. The motivation for this is that, while learners who exhibit *keywording* behavior may enjoy some success in accessing the domain content, they are not taking full advantage of the capabilities of the interface, including the option to delve down deeper into topics, change topics, or ask analytical questions.

To measure potential changes in behavior, a lexical analysis of the learner's input before and after the intervention is examined using the LIWC dataset. A paired-samples T-test was carried out to test for differences between *social presence* and other factors measured before and after the intervention. When we reviewed past

Table 4 LIWC variables for chat log analysis

LIWC Variable	Description
Per-pronoun	Personal pronouns
Pos-emotion	Positive emotion
Neg-emotion	Negative emotion
Social	Social words (collection)
Cog-proc	Cognitive processes (collection)
Perception	Perceptual processes (collection)
Bio	Biological processes (collection)
Relativity	Words describing relationships
Focus-past	Words associated with the past
Focus-present	Words associated with the present
Analytic	Analytical thinking (Pennebaker et al. 2014)
WPS	Words per sentence
Six-letter	Words> 6 letters

conversational logs to identify conversational behaviours, we found two common traits associated with increased conversational engagement are the length of the sentences (WPS) and greater use of 6 letter words or longer (6ltr). The latest version of LIWC also provides a variable called "analytic", a summary variable which indicates use of analytical words, based on research carried out by the authors of LIWC (Pennebaker et al. 2014).

Social presence measures are also expected to increase with conversational engagement because the higher the degree to which the learner views the CA as an intelligent presence, the more they are expected to converse with it in a human-like way. A low social presence rating would suggest that the learner does not differentiate the CA from a data base query application or search engine.

Table 5 shows no significant changes, except for cognitive processes (Pair 5), which increased (not shown). There is no evidence to reject the null hypothesis that there is no difference between the two sets of data.

Intervention 3 – 'Moreing'

As with Intervention 2, the intervention for *morer* behaviour is expected to result in a slight modification to the way the learner interacts with the CA. Specifically, it encourages the learner to add some conversational acts to drive how the content is delivered rather than relying on the systematic, ordered output of the narratives of each topic. The justification for this is that it involves a higher cognitive process to consider different braches in the structure of the topics. *Morer* behaviour, the simple repetition of the same backchannel word, such as "okay", is not very different from simply reading a text book, document, or web content that does not contain hyperlinks. If this is the preferred way to receive information about a topic, there is little incentive for the learner to use this tool.

Table 5 LIWC social presence measures before/after intervention 2

Paired Samples Test

		Paired Differences				Sig (2-tailed)
	LIWC Variable	Mean	Std Dev	t	df	
Pair 1	Per-pronoun	−5.0529	8.3544	−1.600	6	.161
Pair 2	Pos-emotion	−2.1100	7.9004	−.707	6	.506
Pair 3	Neg-emotion	2.3671	6.6544	.941	6	.383
Pair 4	Social	−1.1957	7.1964	−.440	6	.676
Pair 5	Cog-proc	4.4886	4.0608	2.924	6	.026
Pair 6	Perception	.6500	1.3737	1.252	6	.257
Pair 7	Bio	1.6300	3.8934	1.108	6	.310
Pair 8	Relativity	−1.0700	8.9956	−.315	6	.764
Pair 9	Focus-past	−.3600	2.6206	−.363	6	.729
Pair 10	Focus-present	−.7757	8.6712	−.237	6	.821
Pair 11	Authentic	−5.5986	26.3346	−.562	6	.594
Pair 12	WPS	.3614	.7926	1.206	6	.273
Pair 13	Six-letter	1.3400	12.3409	.287	6	.784

The same LIWC variables as those for Intervention 2 were used to detect if the intervention was successful in encouraging the desired change in conversational behavior. Again, paired-samples T-test did not reveal statistically significant differences between social presence and other metrics collected before and after the intervention (Table 6), except for personal pronouns, which showed a significant increase (means not shown, $p = .004$). There is insufficient evidence to reject the null hypothesis.

Discussion and further analysis

LIWC analysis did not find support for the hypothesis that the interventions for keyworders and morers would have a positive influence on social presence, as a dimension of conversational engagement. Unfortunately, there were fewer cases of each (8 morers, 7 keyworders) which made statistical analysis challenging due to small sample size. One possible explanation for the relatively small number of cases of these types of behaviour is that the participants are psychology learners who are taking part in a study for learning about research methods. As such, they may be more inclined to take the task seriously and attempt to ask meaningful questions. This is more in line with the profile of tryer behaviour.

There are a few explanations for why no changes were detected in learner behaviour because of Intervention 2 and 3. These are:

- The interventions were not effective and changes did not take place
- The interventions were inappropriately applied, i.e. *keyworder* and *morer* detection was inaccurate
- The LIWC metrics selected were not sufficiently sensitive to detect changes, particularly given the small sample sizes for these two intervention types

A manual examination of the log files was carried out to confirm whether or not changes in learner behaviour were observed, but not detected, as a result of the

Table 6 LIWC social presence measures before/after intervention 3

Paired Samples Test

		Paired Differences				Sig (2-tailed)
	LIWC Variable	Mean	Std Dev	t	df	
Pair 1	Per-pronoun	−8.3888	5.5038	−4.311	7	.004
Pair 2	Pos-emotion	−1.8563	4.8861	−1.075	7	.318
Pair 3	Neg-emotion	1.9150	5.5921	.969	7	.365
Pair 4	Social	−2.7788	10.9271	−.719	7	.495
Pair 5	Cog-proc	−1.0063	10.1694	−.280	7	.788
Pair 6	Perception	.6825	.8967	2.153	7	.068
Pair 7	Bio	1.0650	2.0265	1.486	7	.181
Pair 8	Relativity	2.1888	12.8729	.481	7	.645
Pair 9	Focus-past	.3138	5.8913	.151	7	.885
Pair 10	Focus-present	1.6538	6.7253	.696	7	.509
Pair 11	Authentic	−28.1250	54.2775	−1.466	7	.186
Pair 12	WPS	−.6075	2.0909	−.822	7	.438
Pair 13	Six-letter	.5163	7.7920	.187	7	.857

interventions. For each participant who received either Intervention 2 or Intervention 3, a simple qualitative analysis was carried out to

- identify the type of conversational behaviour that preceded the intervention
- verify whether the learner's conversational behaviour justified the intervention
- determine if the intervention resulted in an appropriate change in behaviour

The results of the qualitative analysis were generally encouraging. Of those participants receiving Intervention 2, all but one where judged to be showing *keywording* behaviour, meaning that the intervention was justified. 5 (of 7) participants responded to the intervention by attempting at least a few full, though sometimes short, sentences. 1 of these reverted back to non-conversational behavior. This appeared to be a result of poor performance on the part of the CA. This suggests that the intervention is successful in motivating the learner to explore the option to converse with the CA, but still requires the CA to do its part by rewarding the learner with an improved experience, i.e. provide domain information. Similar results were found with participants who received Intervention 3. All learners exhibited *moreing* behaviour, justifying the intervention. Seven out of 8 cases resulted in a change of behaviour. One case was inconclusive at it occurred at the end of the conversation. Of those that did switch to full sentences after the intervention, 2 reverted back to *moreing* after attempts were met with no-match responses. The fact that we observe the desired behavior change after Interventions 2 and 3, but our qualitative tests do not show this, suggests that the tests are not sensitive enough, or may be completely ineffective as a way to measure conversational engagement. We will require more data to verify which case is true. Because the behaviour change resulting from the interventions was temporary in some cases, it also appears that the effectiveness of the interventions may be limited by other factors, such as the performance of the CA.

Conclusions and future work

We have proposed an approach to enabling the capability of Fredubot in intelligent intervention and strategic pedagogical design to improve learner engagement in learning through chatting with the conversation agent. We have developed a set of algorithms that classify learner behavior types which indicate low conversational engagement for our CA, Freudbot. Agents are used to detect these learner behaviours through real-time monitoring and analysis of the ongoing transcript of the interaction, and work with the CA to initiate appropriate interventions. Conversational interventions were designed and applied according to behaviour types. This research implemented a method of incorporating intelligence and decision making support in using interventions into an existing CA through developing an Intervention agent in the system. In the future, additional attributes could be explored using lexical analysis tools such as LIWC to preprocess the utterances in real-time. Also, O'Shea's work with *dialogue act identification techniques* and *semantic processing* to determine sentence similarity (O'Shea 2012) suggests the potential for a new agent for processing adjacency pairs for testing similarity between CA output and learner input to judge conversational engagement.

Although not all interventions were effective in changing behaviour, as a group, interventions were associated with a significantly greater number of exchanges compared to uncategorized learners exposed to no interventions. Although the focus of this research was on interventions to improve engagement, the findings show more generally that adaptive interventions that reflect intelligent detection and decision making may be critical in the successful design of CA as pedagogical agents. The multiple agent framework is ideally suited to realize these types of adaptive interventions based not only on conversation but on other affective channels of communication.

Acknowledgements
Not applicable.

Funding
Not applicable.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Authors' contributions
MP performed all research and writing of the manuscript. FL and BH provided supervision of the research, insight and editing of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Athabasca University Research Ethics Board reviewed the research and approved on ethical grounds on June 22, 2016.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]School of Computing and Information Systems, Athabasca University, Athabasca, Canada. [2]Faculty of Humanities and Social Sciences, Athabasca University, Athabasca, Canada.

Received: 13 July 2018 Accepted: 2 November 2018
Published online: 20 November 2018

References

S. Asteriadis, K. Karpouzis, S. Kollias, in *Human-Computer Interaction. New Trends*, ed. by J. A. Jacko. Feature extraction and selection for inferring user engagement in an HCI environment (Springer, Berlin Heidelberg, 2009), pp. 22–29 https://link.springer.com/chapter/10.1007/978-3-642-02574-7_3

C. Becker, S. Kopp, I. Wachsmuth, *Why emotions should be integrated into conversational agents*, Conversational Informatics: An Engineering Approach (2007), pp. 49–68

F. Bellotti, R. Berta, A. De Gloria, E. Lavagnino, in *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*. Towards a conversational agent architecture to favor knowledge discovery in serious games (ACM, New York, 2011), pp. 17:1–17:7 https://doi.org/10.1145/2071423.2071444

J. Cassell, Embodied conversational agents: Representation and intelligence in user interfaces. AI Mag. **22**(4), 67 (2001)

G. Castellano, A. Pereira, I. Leite, A. Paiva, P.W. McOwan, in *Proceedings of the 2009 International Conference on Multimodal Interfaces*. Detecting user engagement with a robot companion using task and social interaction-based features (ACM, New York, 2009), pp. 119–126 https://doi.org/10.1145/1647314.1647336

J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2 edition. (Hillsdale, Routledge, 1988).

S. D'Mello, S. Craig, A. Witherspoon, B. McDaniel, A. Graesser, Automatic detection of learner's affect from conversational cues. User Model. User-Adap. Inter. **18**(1), 45–80 (2008) https://doi.org/10.1007/s11257-007-9037-6

D.R. Danforth, M. Procter, R. Chen, M. Johnson, R. Heller, Development of virtual patient simulations for medical education. J Virtual Worlds Res **2**(2) (2009) https://journals.tdl.org/jvwr/index.php/jvwr/article/view/707

A.C. Graesser, N.K. Person, X. Hu, Improving comprehension through discourse processing. New Dir. Teach. Learn. **2002**(89), 33–44 (2002) https://doi.org/10.1002/tl.45

B. Heller, *Conversational agents as historical figures: Individual differences and perceptions of agent and social presence*, Presented at the EdMedia: World conference on educational media and Technology, vol 2016 (2016), pp. 1368–1374 https://www.learntechlib.org/p/173133/

B. Heller, M. Procter, *Embodied and embedded intelligence: Actor Agents on Virtual Stages. Intelligent and Adaptive Learning Systems: Technology Enhanced Support for Learners and Teachers* (2011), pp. 280–290

B. Heller, & Procter, M. Conversational Agents in Virtual Worlds: Immersion and the Conversational Record. In S. Gregory, P. Jerry, & N. Tavares-Jones (Eds.), At the Edge of the Rift. (Freeland, Oxfordshire, United Kingdom: Inter-Disciplinary Press, 2014), pp. 25–36

B. Heller, M. Procter, C. Rose, in *Learning in Virtual Worlds*, ed. by S. Gregory, M. J. W. Lee, B. Dalgarno, B. Tynan. Conversational agents in second life (Athabasca University Press, 2016), pp. 153–166

R. Heller, M. Procter, Animated pedagogical agents: The effect of visual information on a historical figure application. Int J Web-Based Learn Teach Technolo **4**(1), 54–65 (2009) https://doi.org/10.4018/jwltt.2009010104

W.L. Johnson, J.W. Rickel, J.C. Lester, et al., Animated pedagogical agents: Face-to-face interaction in interactive learning environments. Int. J. Artif. Intell. Educ. **11**(1), 47–78 (2000)

A. Kapoor, R.W. Picard, in *Proceedings of the 13th Annual ACM International Conference on Multimedia*. Multimodal affect recognition in learning environments (ACM, New York, 2005), pp. 677–682 https://doi.org/10.1145/1101149.1101300

R. Kumar, C.P. Rosé, Architecture for building conversational agents that support collaborative learning. IEEE Trans. Learn. Technol. **4**(1), 21–34 (2011) https://doi.org/10.1109/TLT.2010.41

F. Mairesse, M.A. Walker, M.R. Mehl, R.K. Moore, Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Intell. Res. **30**(1), 457–500 (2007)

G. McClure, M. Chang, F. Lin, in *Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on*. MAS controlled NPCs in 3D virtual learning environment (2013), pp. 1026–1033 IEEE. Retrieved from https://ieeexplore.ieee.org/document/6727316

Y.I. Nakano, R. Ishii, in *Proceedings of the 15th International Conference on Intelligent User Interfaces*. Estimating User's engagement from eye-gaze behaviors in human-agent conversations (ACM, New York, 2010), pp. 139–148 https://doi.org/10.1145/1719970.1719990

N. Novielli, HMM modeling of user engagement in advice-giving dialogues. J Multimodal User Interfaces **3**(1–2), 131–140 (2010) https://doi.org/10.1007/s12193-009-0026-4

K. O'Shea, An approach to conversational agent design using semantic sentence similarity. Appl. Intell. **37**(4), 558–568 (2012) https://doi.org/10.1007/s10489-012-0349-9

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. Retrieved from https://repositories.lib.utexas.edu/handle/2152/31333

J.W. Pennebaker, C.K. Chung, J. Frazee, G.M. Lavergne, D.I. Beaver, When small words foretell academic success: The case of college admissions essays. PLoS One **9**(12), e115844 (2014) https://doi.org/10.1371/journal.pone.0115844

D. Perez-Marin, I. Pascual-Nieto, in *Conversational Agents and Natural Language Interaction*. Future Trends for Conversational Agents (2011), pp. 395–400 IGI Global. https://www.igi-global.com/gateway/chapter/54648

M. Procter, F. Lin, R. Heller, in *Advances in Artificial Intelligence*. Improving Conversation Engagement through Data-Driven Agent Behavior Modification (Springer, Cham, 2016), pp. 270–275 https://doi.org/10.1007/978-3-319-34111-8_33

D. Szafir, B. Mutlu, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Pay attention!: Designing adaptive agents that monitor and improve user engagement (ACM, New York, 2012), pp. 11–20 https://doi.org/10.1145/2207676.2207679

P.D. Turney, Y. Neuman, D. Assaf, Y. Cohen, in *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*. Literal and metaphorical sense identification through concrete and abstract context (2011), pp. 680–690 Retrieved from http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=18533382

G. Veletsianos, G.S. Russell, What do learners and pedagogical agents discuss when given opportunities for open-ended dialogue? J. Educ. Comput. Res. **48**(3), 381–401 (2013) https://doi.org/10.2190/EC.48.3.e

Wen, M., Yang, D., & Rose, C. P. (2014). Linguistic reflections of student engagement in massive open online courses. In Eighth International AAAI Conference on Weblogs and Social Media Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8057

A. Wise, J. Chang, T. Duffy, R.D. Valle, The effects of teacher social presence on student satisfaction, engagement, and learning. J. Educ. Comput. Res. **31**(3), 247–271 (2004) https://doi.org/10.2190/V0LB-1M37-RNR8-Y2U1

Q. Xu, L. Li, G. Wang, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Designing engagement-aware agents for multiparty conversations (ACM, New York, 2013), pp. 2233–2242 https://doi.org/10.1145/2470654.2481308

K. Yamashita, H. Kubota, T. Nishida, Designing conversational agents: Effect of conversational form on our comprehension. AI & Soc. **20**(2), 125–137 (2005) https://doi.org/10.1007/s00146-005-0011-8