Modelling e-learner comprehension within a conversational intelligent tutoring system

Abstract. Conversational Intelligent Tutoring Systems (CITS) are agent based e-learning systems which deliver tutorial content through discussion, asking and answering questions, identifying gaps in knowledge and providing feedback in natural language. Personalisation and adaptation for CITS are current research focuses in the field. Classroom studies have shown that experienced human tutors automatically, through experience, estimate a learner's level of subject comprehension during interactions and modify lesson content, activities and pedagogy in response. This paper introduces Hendrix 2.0, a novel CITS capable of classifying e-learner comprehension in real-time from webcam images. Hendrix 2.0 integrates a novel image processing and machine learning algorithm, COMPASS, that rapidly detects a broad range of non-verbal behaviours, producing a time-series of comprehension estimates on a scale from -1.0 to +1.0. This paper reports an empirical study of comprehension classification accuracy, during which 51 students at Manchester Metropolitan University undertook conversational tutoring with Hendrix 2.0. The authors evaluate the accuracy of strong comprehension and strong non-comprehension classifications, during conversational questioning. The results show that the COMPASS comprehension classifier achieved normalised classification accuracy of 75%.

Keywords. Conversational intelligent tutoring systems, comprehension assessment, machine learning

1. Introduction

E-learning is changing the way in which students engage with learning materials. From video lectures to web-based content distribution and quizzes, the process of learning is shifting from human- to software-directed learning. Whilst much of the advancement in e-learning has focused on increasing the availability of information through digital technology, it has often neglected the important role that personalised tuition plays in supporting learning [1] and developing higher order skills.

Conversational Intelligent Tutoring Systems (CITS) [2] are a virtual assistant technology that place pedagogy at the heart of content delivery. CITS use natural language to discuss concepts with the learner, breaking down the tuition material into conversational interactions, asking and answering questions, identifying gaps in knowledge, and providing contextual feedback and corrective interventions. CITS

have been shown to outperform self-directed learning and both digital and traditional book-based learning [3, 4]. To further improve the effectiveness of CITS tuition, research has turned to adaptive algorithms. Adaptive CITS personalise the learning experience for the student by, for example, changing content and pedagogy in response to learning style [5, 6] or emotional valence [7, 8].

Classroom studies [9–11] have shown that experienced human tutors can accurately estimate a learner's level of subject comprehension based on non-verbal behaviour. Non-verbal behaviour (NVB) is any non-verbalised communicative behaviour including gestures, facial expressions, facial actions, physiological, chemical and audible information. Cutting edge research [12] suggests it is possible to model comprehension automatically by computer analysis of non-verbal behaviour. Buckingham et al. [12] demonstrate that learner NVB can be used to model learner comprehension expressed during pre-recorded human to human interviews, conducted in a strictly controlled environment.

While the research [12, 13] indicates that learner comprehension levels could be used as a feedback mechanism for adaptation within a CITS, doing so in a live real-world environment presents a number of novel challenges:

a) Analysing live video in near real-time to track NVB in an uncontrolled environment (varying lighting, camera position),
b) Accurately modelling comprehension from NVB during conversation with a virtual tutor,
c) Producing accurate comprehension modelling without invasive body-attached sensors or prohibitively expensive specialised equipment.

In this paper the authors present an empirical study of comprehension classification integrated into a conversational intelligent tutoring system. The research aims to demonstrate that real-time classification of e-learner comprehension is a viable feedback mechanism for adaptation within a CITS. In a study conducted at Manchester Metropolitan University, 51 higher education students undertook a short course of tuition using a novel CITS, called Hendrix 2.0. Hendrix 2.0 has been designed to tutor computer programming at an undergraduate level. During the tutorial, Hendrix 2.0 automatically modelled learner comprehension in real-time using a novel machine learning based comprehension assessment and scoring system, now called COMPASS [14].

To evaluate the effectiveness of COMPASS real-time comprehension modelling within a CITS, the authors present classification accuracy results for binary 'strong comprehension' and 'strong non-comprehension' classifications. COMPASS achieved normalised classification accuracy of 75%. The results demonstrate that the comprehension modelling algorithm overcomes novel challenges, accurately modelling e-learner NVB during conversation with a virtual tutor, analysing NVB in near real-time in an uncontrolled classroom environment, and producing accurate comprehension classifications.

This paper is organised as follows: Section 2 presents an overview of related research and prior work on conversational intelligent tutoring systems (2.1), interpretation of learner non-verbal behaviour (2.2) and automatic comprehension classification (2.3). Section 3 introduces Hendrix 2.0, a novel CITS, and describes the

question-answer and comprehension classification processes. A study of COMPASS comprehension classification accuracy within Hendrix 2.0 is presented in Section 4, with results and discussion (4.3). Section 5 presents the authors' conclusions and Section 6 presents the intended future work for this research project.

2. Related work

2.1. Conversational intelligent tutoring systems

Conversational Intelligent Tutoring Systems (CITS) [2] are a virtual assistant technology that place pedagogy at the heart of content delivery. CITS use natural language to discuss concepts with the learner, breaking down the tuition material into conversational interactions, asking and answering questions, identifying gaps in knowledge and providing contextual feedback and corrective interventions.

In prior research [15] the authors designed and developed a conversational intelligent tutoring system (CITS) named Hendrix. The Hendrix CITS chats with a learner using written natural language, via a software interface. Hendrix is a goal-oriented conversational agent, designed to follow a scripted learning scaffold which is embedded within a graph of concepts. Hendrix is able to find connections between concepts in the graph and is then able to ask the learner questions, appraise the learner's knowledge and provide contextually relevant feedback. Hendrix asks the learner to demonstrate understanding by answering open questions. Hendrix can then appraise the correctness of the response by matching the learner's discursive response to a bank of pre-defined exemplar answer patterns. While Hendrix can interpret, match and search information in a sophisticated way, the pilot version of Hendrix is naïve to the learner.

The tutor's role in mediating the learning experience is personal, intimate, and uniquely crafted to the needs of the learner in question. Personalisation algorithms for CITS have become a focus of research in the field. Adaptive CITS use contextual personal information about the learner, their preferences, behaviour, emotions, or cognitive functions, to adapt the conversational tutorial content or system behaviour to meet the needs of the individual learner. Latham et al. [5, 6] demonstrated the effectiveness of adapting to learning style; others [7, 16, 17, 8, 18] have focused on the role of affect – or emotional valence – in mediating the tuition.

In this paper the authors introduce Hendrix 2.0, a novel CITS capable of modelling learner comprehension in real-time by observing learner non-verbal behaviour.

2.2. Interpreting the non-verbal behaviour of learners

Non-verbal behaviour (NVB) is any non-verbalised communicative behaviour, including movement, gesture, facial expression, sound, and chemical and physiological signals. Facial actions and expressions have been used to model learner emotional states such as boredom, confusion and frustration [7, 8, 18, 19].

Classroom studies [9–11] have shown that expert human tutors are able to recognise learner comprehension and non-comprehension states by observing the learner's non-verbal behaviour. By recognising and responding to comprehension indicative non-verbal behaviour, expert human tutors are able to prevent loss of motivation, feelings of hopelessness, frustration and boredom during complex learning, adapting tuition content and technique in response to non-comprehension impasse as it occurs. While affect cannot be dismissed as a useful feedback channel for an intelligent e-learning system, the flow of cognitive and affective states suggests that detectable affect may occur too late in the learning process to help prevent repeated impasse and loss of motivation. Hendrix 2.0 overcomes this problem by modelling learner comprehension in real-time, allowing the virtual tutor to identify and respond to points of impasse as they occur.

2.3. Comprehension classification by automata

FATHOM [12] is a comprehension classification system which analyses learner non-verbal behaviour to automatically classify comprehension during tutorial questioning. The work demonstrates that video recordings of learners responding verbally to a human tutor can be computationally analysed to produce accurate classifications of learner comprehension levels during verbal student-tutor interactions.

Hendrix 2.0 integrates the real-time comprehension assessment and scoring system, COMPASS [14], to monitor learner comprehension during conversational tutoring. COMPASS is a novel artificial neural network (ANN) based classifier which is able to analyse live image stream data to survey observable learner non-verbal behaviour rapidly. COMPASS uses a combination of computer vision [20] and machine learning [21] techniques to produce a comprehension estimate for each one second of video footage. The system tracks 37 non-verbal behaviours, including head movement and rotation, eye gaze direction, blink rate, skin tone change and 6 meta-data variables, including gender and ethnicity. Observed behaviours for each one second of image stream data are summarised to produce a 43 variable cumulative behavioural feature vector (CBFV), a numeric vector representing the average behaviour of the learner over the one second time window. The CBFV is classified using a multilayer perceptron (MLP) network [21], outputting a single comprehension estimate on the scale -1.0 to +1.0, where -1.0 is non-comprehension and +1.0 is comprehension. Binary classification is achieved by applying a logistic threshold.

In prior work, the COMPASS comprehension classifier was optimised and trained using pre-tagged recordings of students answering multiple choice questions [14]. Each recording was tagged as either comprehension or non-comprehension, depending on the correctness of the answer selected by the learner. The classifier achieved test normalised classification accuracy of 74%.

3. Hendrix 2.0

Hendrix 2.0 integrates two novel artificially intelligent systems: a novel conversational intelligent tutoring system, based on Hendrix 1.0 [15], and a novel comprehension classification and scoring system, now called COMPASS [14].

Fig. 1. Hendrix 2.0 chat interface window (left) and real-time comprehension monitor window (right)

Hendrix 2.0 interacts with the learner via a chat interface, as shown in figure 1. During interactions, Hendrix 2.0 uses COMPASS to model the learner's comprehension and produce a time-series of comprehension estimates and classifications (figure 1).

Hendrix 2.0 structures conversational tutorial content by searching a graph of pre-defined subject knowledge which includes hierarchically structured concepts, definitions, examples, code samples, questions and more. During the conversation, Hendrix 2.0 will appraise the learner's understanding of the subject by asking the learner to explain concepts, answer questions or write basic programming code. Hendrix 2.0 can ask both simple questions requiring just a single word answer, for example 'true' or 'false', or complex questions requiring multiple words, phrases, mathematics or programming code.

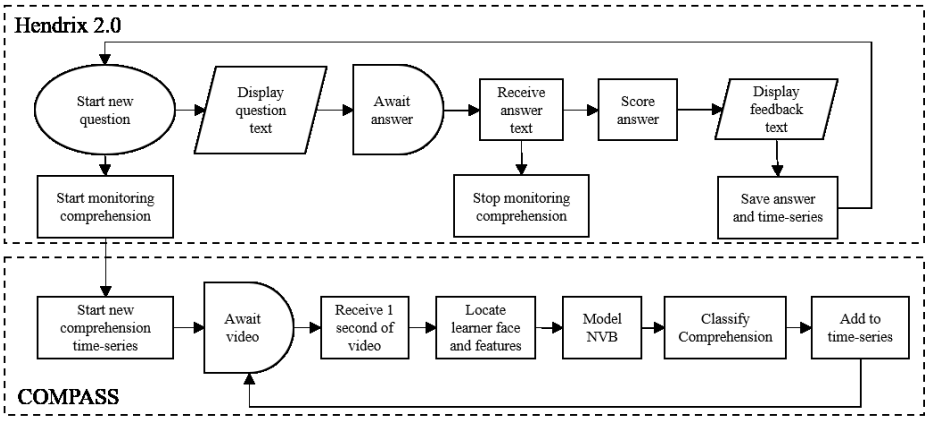

Fig. 2. Question, answer and comprehension monitoring process for Hendrix 2.0 with COMPASS integration

Figure 2 shows the process by which Hendrix 2.0 asks, scores and saves learner's answers. When a new question is loaded, Hendrix 2.0 simultaneously displays the question text and begins monitoring the learner's comprehension. COMPASS creates a new time-series and awaits video data to analyse. Hendrix 2.0 captures video footage from a front-facing webcam and passes it to COMPASS one second at a time. Each one second of video footage is analysed to produce a comprehension estimate, which is added to the time-series for the answer response period. Once an answer is submitted by the learner, comprehension monitoring is stopped, the answer is automatically scored by Hendrix 2.0 and both the answer and the comprehension time-series are saved.

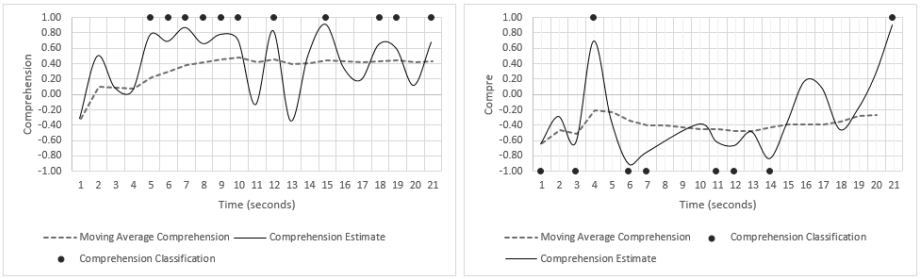

Fig. 3. COMPASS learner comprehension modelling for tutorial answers scoring 100% correct (left) and 0% correct (right)

Figure 3 shows two exemplar comprehension time-series, generated during the study. The raw classifier output is shown as 'Comprehension', a continuous stream of values between -1.0 and 1.0 where -1.0 is strong non-comprehension and +1.0 is strong comprehension. 'Classification' dots show binary comprehension classification, using a threshold of ±0.6. A binary classification is made when the network output is either greater than or equal to 0.6 (*'strong comprehension'*), or when the network output is less than or equal to -0.6 (*'strong non-comprehension'*).

4. Study: automatic classification of e-learner comprehension during on-screen conversational interactions

The aim of this experiment is to investigate whether COMPASS can be integrated into a conversational intelligent tutoring system, such as Hendrix 2.0, and whether the COMPASS comprehension classifier can accurately classify e-learner comprehension in real-time, from non-verbal behaviour, during conversational interactions.

4.1. Participants

The participant group consisted of 51 students from the Faculty of Science & Engineering at Manchester Metropolitan University. The gender demographics of the participant group were 80% male and 20% female. White European (or Caucasian) was the largest ethnic group, at just over 50%, with Asian (or British Asian) the second largest at 35%.

4.2. Method

In a study conducted at Manchester Metropolitan University, participants undertook a tutorial on computer programming using the Hendrix 2.0 CITS (figure 1). For each answer given by a learner, Hendrix 2.0 automatically assigned a score between 0 and 100%. During each answer period, Hendrix 2.0 programmatically passed a stream of web camera image data to the COMPASS [14] API for real-time comprehension classification (figure 2).

The 51 participants answered a total of 1,269 questions, answering on average 26 questions each. Across all answers, COMPASS classified 14,931 seconds non-verbal behaviour, and produced 7,027 binary comprehension classifications. The accuracy of comprehension classification is evaluated by calculating the true positive (predicted comprehension) and true negative (predicted non-comprehension) percentages for classifications made during answer periods scoring 50% or higher (observed comprehension) and periods scoring 25% or lower (observed non-comprehension).

4.3. Results and discussion

Research [22] has found that in practical application culture, social norms, ethnicity, and gender all play an important role in mediating non-verbal behaviour. In line with the findings of Rothwell et al. [22], and the broader literature on cross-cultural non-verbal behaviour [23], separate classifiers may need to be trained for each demographic group to optimise performance. The results presented in this paper focus on the largest demographic subset of participants – white European males – who account for 58% of the data collected during the experiment.

For binary classification, true non-comprehension periods are defined as response periods for answers scoring 25% or lower. True comprehension periods are defined as response periods scoring 50% or higher.

Table 1. Classification accuracy for white males

Counts			Classification accuracy (%)		
Answers	Classifications	Threshold	Non-comp	Comp	Norm
532	3313	±0.6	54.09	68.74	60.64
	1926	±0.8	64.05	67.89	65.73

Table 1 shows the comprehension classifier accuracy for all white male response periods. At threshold ±0.6 3313 classifications are made during the 532 answer response periods, with a normalised classification accuracy (NCA) of 60%. Increasing the threshold to ±0.8 reduces the number of classifications made by 40%, but increases NCA by 5%. While the results in table 1 indicate that it is possible to detect comprehension and non-comprehension, classification accuracy is weak.

Table 2. Classification accuracy for white males by question difficulty

	Counts			Classification accuracy (%)		
Question difficulty	Answers	Classifications	Threshold	Non-comp	Comp	Norm
Simple	200	986	±0.6	35.48	52.90	42.49
		574	±0.8	36.56	50.79	42.86
Complex	332	2327	±0.6	62.91	74.54	68.33
		1352	±0.8	75.59	75.25	75.44

Table 2 shows comprehension classification broken down by question complexity. Classification accuracy at chance levels for simple questions may be explained by guessing. Simple questions require only a single word answer, such as '*true*' or '*false*', allowing a learner to guess the answer easily. Guessing behaviour results in '*strong non-comprehension'* behaviour becoming associated with a high answer score, reducing the evident accuracy of the classifier. Complex questions, which cannot be guessed easily require learners to formulate complex multi-part answers containing multiple keywords, phrases, mathematics or programming code, which can contain multiple conceptual elements. For complex questions the classifier achieves normalised classification accuracy of between 68% and 75%, similar to accuracy during training and testing of the classifier [14].

The results demonstrate that the COMPASS classifier can, with accuracy, detect and classify '*strong comprehension*' and '*strong non-comprehension*' indicative behaviours during conversational interactions with a conversational intelligent tutoring system, given the limitations highlighted. The results also highlight the consideration which should be given to the type of question being addressed to the learner, whether guessing behaviour is likely and whether the question challenges the learner sufficiently to evoke strong comprehension or non-comprehension behaviour.

5. Conclusions

In this paper the authors have presented a review of literature demonstrating the novelty of, and motivation for, the development of a method for real-time, automatic, modelling of e-learner comprehension by computational analysis of non-verbal behaviour. The authors have introduced a novel conversational intelligent tutoring system (CITS), named Hendrix 2.0, which is capable of modelling comprehension in real-time using a comprehension assessment and scoring system, named COMPASS.

The authors have presented an experiment in which real learners undertake conversational tutoring using Hendrix 2.0, and their scoring tutorial answers and COMPASS comprehension time-series' are saved. The authors present results and discussion of classification accuracy and find that the COMPASS classifier is able to achieve 75% classification accuracy on *'complex'* questions. The research presented here demonstrates that comprehension classification by computational analysis of learner non-verbal behaviour is a viable feedback mechanism for an adaptive conversational intelligent tutoring system, such as Hendrix 2.0. The results do highlight limitations and considerations: separate classifiers should be trained for each gender and ethnicity, and a separate classifier may also be needed for *'simple'* questions, to identify and model behaviours associated with guessing.

6. Future work

In future work the authors will investigate whether CITS adaptation based on '*strong non-comprehension*' classification improves a learner's tutorial score and post-tutorial test score performance, when compared to a control group.

References

1. Salmon, G.: Flying not flapping: a strategic framework for e-learning and pedagogical innovation in higher education institutions. ALT-J. 13, 201–218 (2005).
2. VanLehn, K.: The behavior of tutoring systems. Int. J. Artif. Intell. Educ. 16, 227–265 (2006).
3. VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., Rosé, C.P.: When are tutorial dialogues more effective than reading? Cogn. Sci. 31, 3–62 (2007).
4. VanLehn, K.: The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. Educ. Psychol. 46, 197–221 (2011).
5. Latham, A., Crockett, K., McLean, D.: An adaptation algorithm for an intelligent natural language tutoring system. Comput. Educ. 71, 97–110 (2014).
6. Latham, A., Crockett, K., McLean, D., Edmonds, B.: A conversational intelligent tutoring system to automatically predict learning styles. Comput. Educ. 59, 95–109 (2012).
7. D'Mello, K.S., Craig, S.D., Gholson, B., Franklin, S., Picard, R., Graesser, A.C.: Integrating affect sensors in an intelligent tutoring system. In: Affective Interactions: The Computer in the Affective Loop Workshop at. pp. 7–13 (2005).
8. Lin, H.-C.K., Wu, C.-H., Hsueh, Y.-P.: The influence of using affective tutoring system in accounting remedial instruction on learning performance and usability. Comput. Hum. Behav. 41, 514–522 (2014).

9. Alibali, M.W., Flevares, L.M., Goldin-Meadow, S.: Assessing knowledge conveyed in gesture: Do teachers have the upper hand? J. Educ. Psychol. 89, 183–193 (1997).
10. Machida, S.: Teacher Accuracy in Decoding Nonverbal Indicants of.pdf. J. Educ. Psychol. 78, 454–464 (1986).
11. Webb, J.M., Diana, E.M., Luft, P., Brooks, E.W., Brennan, E.L.: Influence of pedagogical expertise and feedback on assessing student comprehension from nonverbal behavior. J. Educ. Res. 91, 89–97 (1997).
12. Buckingham, F.J., Crockett, K.A., Bandar, Z.A., O'Shea, J.D.: FATHOM: A neural network-based non-verbal human comprehension detection system for learning environments. In: Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on. pp. 403–409. IEEE (2014).
13. Buckingham, F.J., Crockett, K.A., Bandar, Z.A., O'Shea, J.D., MacQueen, K.M., Chen, M.: Measuring human comprehension from nonverbal behaviour using artificial neural networks. In: Neural Networks (IJCNN), The 2012 International Joint Conference on. pp. 1–8. IEEE (2012).
14. Holmes, M., Latham, A., Crockett, K., O'Shea, J.D.: Real-time comprehension classification with artificial neural networks: decoding e-Learner non-verbal behaviour, (2016).
15. Holmes, M., Latham, A., Crockett, K., O'Shea, J.D., Lewin, C.: Hendrix: A conversational intelligent tutoring system for Java programming. Presented at the UK workshops on Computational Intelligence , University of Exeter (2015).
16. Whitehill, J., Bartlett, M., Movellan, J.: Automatic facial expression recognition for intelligent tutoring systems. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on. pp. 1–6. IEEE (2008).
17. Calvo, R.A., D'Mello, S.: Frontiers of affect-aware learning technologies. IEEE Intell. Syst. 27, 86–89 (2012).
18. Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., Zhao, W.: Automatic Detection of Learning-Centered Affective States in the Wild. Presented at the (2015).
19. Whitehill, J., Serpell, Z., Foster, A., Lin, Y.-C., Pearson, B., Bartlett, M., Movellan, J.: Towards an optimal affect-sensitive instructional system of cognitive skills. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on. pp. 20–25. IEEE (2011).
20. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. 57, 137–154 (2004).
21. Haykin, S.: Neural Networks: A Comprehensive Foundation. Macmillan (1994).
22. Rothwell, J., Bandar, Z., O'Shea, J., McLean, D.: Silent talker: a new computer-based system for the analysis of facial cues to deception. Appl. Cogn. Psychol. 20, 757–777 (2006).
23. Bond, C.F., Omar, A., Mahmoud, A., Bonser, R.N.: Lie detection across cultures. J. Nonverbal Behav. 14, 189–204 (1990).