Leveraging Hidden Dialogue State to Select Tutorial Moves

Kristy Elizabeth Boyer, Robert Phillips,
Eun Young Ha, Michael Wallis,
Mladen Vouk, & James Lester

Department of Computer Science
North Carolina State University
Introduction
Tutorial Dialogue Systems

- Geometry (Aleven et al., 2004)
- Circulatory System (Evens & Michael, 2006)
- Physics (Graesser et al., 2005; Jordan et al., 2006)
- Data Structures (Fossati et al., 2008)
- And many more...
Challenge

Never achieved effectiveness equal to the most effective human tutors

\[ \mu_{\text{control}} \quad \mu_{\text{tutoring}} = \mu_{\text{control}} + 2\sigma_{\text{control}} \]

(Bloom, 1984)
Reasons?

- Need more sophisticated natural language dialogue (Graesser et al., 1994)

AND/OR

- Need to facilitate mastery learning (Van Lehn, 2008)
Data-driven authoring of tutorial dialogue system behavior
Hypothesized Benefits

Data-driven tutorial dialogue systems:
- Reduced development time
- Increased number of hours of instruction for students
- Flexible dialogue strategies
- Reflect approaches of effective human tutors
Goal of This Work

Learn

a data-driven tutorial dialogue strategy

as evidenced by

predicting human tutors’ dialogue acts

within a corpus
Approach

Perform sequence prediction with models that leverage hidden dialogue state
Data-driven dialogue policy creation through:

- Reinforcement learning
- Direct corpus-based extraction
Reinforcement Learning

- Markov assumption (Levin et al. 2000)
- Reinforcement learning (Frampton & Lemon, 2009)
- Challenges include sparse data and large state spaces (Ai et al., 2007; Tetreault & Litman, 2008; Henderson et al., 2008; Heeman, 2007; Young et al., 2009)
- Comparing specific tutorial dialogue tactic choices (Chi et al., 2008)
Corpus-Based Extraction

Assumes that a good dialogue policy is realized in successful human-human dialogues

- Financial domain (Hardy et al., 2006)
- Catalogue ordering (Bangalore et al., 2008)
- Maptask conversational game (Poesio & Mikheev, 1998)
Data-driven exploration

- CIRCSIM-TUTOR (Evens & Michael, 2006)
- ITSPROKE (Forbes-Riley, Rotaru, Litman, & Tetreault, 2007; Forbes-Riley & Litman, 2009)
- KSC-PAL (Kersey, Di Eugenio, Jordan, & Katz, 2009)
Hidden Dialogue State

- Learned by hidden Markov models (Boyer et al., 2009a)
- Qualitative resemblance to tutorial dialogue strategies (Boyer et al., 2009b)
- Hypothesized to boost prediction of tutor moves
Corpus Study

- 48 Human-human tutoring dialogues
- Domain: introductory computer programming
- Task-oriented, separate, parallel task event stream
Corpus Collection

Room #1
Student

Room #2
Tutor

Programming Actions and Typed Dialogue
Typed Dialogue
Remote Tutoring Environment

Tutor (Navigator 08:10:01): No, we've already declared it. We can just use it.
Student (Crver 08:11:10): I'm sorry. I'm still stuck on that last loop we wrote because I feel like that's where my digits should come from.
Tutor (Navigator 08:11:18): That's fine.
Tutor (Navigator 08:11:25): That is where the digits are coming from.
Tutor (Navigator 08:11:32): They're stored in the digits array.
Tutor (Navigator 08:11:41): Does that make sense?
Student (Crver 08:12:25): Sort of, but then how do I add them together? Is it something like digits[1] + digits[2] + digits[3] + ... or is there a different way to approach this since we're using an array?
Student (Crver 08:12:43): That's exactly what we're going to be doing, but we can do it with

Nobody is typing.

/**
 * Extract the individual digits stored in the ZIP code
 * and store their values as private data
 */

private void extractDigits() {

    for (int i = 0; i < 5; i++) {
        digits[i] = (int)zipCode / Math.pow(10, 5 - i))
    }
}

barCode.clearCode();

/**
 * Calculate correction digit and draw it
 * (used by actionPerformed())
 */
Corpus

- 1,468 student utterances and 3,338 tutor utterances
- 3,793 semantic student task actions
- Significant learning gain from pretest to posttest (7% average, $p<0.0001$)
- Annotated with dialogue act tags and task/subtask structure
Dialogue Act Annotation

- Scheme inspired by tags for
  - Conversational speech (Stolcke et al., 2000)
  - Task-oriented dialogue (Core & Allen, 1997)
  - Tutoring (Litman & Forbes-Riley, 2006)

- Inter-rater reliability on 10% of corpus was $\kappa=0.80$. 
# Dialogue Act Tags

<table>
<thead>
<tr>
<th>Dialogue Act Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSESSING</td>
<td>Request for feedback on task or conceptual utterance.</td>
</tr>
<tr>
<td>QUESTION (AQ)</td>
<td></td>
</tr>
<tr>
<td>EXTRA-DOMAIN (EX)</td>
<td>Asides not relevant to the tutoring task.</td>
</tr>
<tr>
<td>GROUNDING (G)</td>
<td>Acknowledgement/thanks.</td>
</tr>
<tr>
<td>LUKEWARM CONTENT FEEDBACK (LCF)</td>
<td>Negative assessment with explanation.</td>
</tr>
<tr>
<td>LUKEWARM FEEDBACK (LF)</td>
<td>Lukewarm assessment of task action or conceptual utterance.</td>
</tr>
<tr>
<td>NEGATIVE CONTENT FEEDBACK (NCF)</td>
<td>Negative assessment with explanation.</td>
</tr>
<tr>
<td>NEGATIVE FEEDBACK (NF)</td>
<td>Negative assessment of task action or conceptual utterance.</td>
</tr>
<tr>
<td>POSITIVE CONTENT FEEDBACK (PCF)</td>
<td>Positive assessment with explanation.</td>
</tr>
<tr>
<td>POSITIVE FEEDBACK (PF)</td>
<td>Positive assessment of task action or conceptual utterance.</td>
</tr>
<tr>
<td>QUESTION (Q)</td>
<td>Task or conceptual question.</td>
</tr>
<tr>
<td>STATEMENT (S)</td>
<td>Task or conceptual assertion.</td>
</tr>
</tbody>
</table>
Task Annotation

- Corpus includes 97,509 keystroke-level student task events
- Manually aggregated into 3,793 subtask actions
- Annotated with task/subtask structure
- Marked for correctness
Subtask Structure Annotation

- Declare New Array
  - double []
  - newTimesArray
  - = new double
  - [times.length]

- Create Parallel Arrays (Subtask 2)

- Construct Conditional
  - Case: value < 5
    - if (times[i]<5)
    - newTimesArray[i] =
    - times[i]+1
  - Case: else
    - else
    - newTimesArray[i] =
    - times[i]
Task Annotation Results

- Subtask annotation (66 leaves)
  - Inter-rater reliability study on 20% of corpus
  - Simple kappa = 0.58
  - Weighted kappa = 0.86

- Correctness annotation (4 tags)
  - CORRECT, BUGGY, INCOMPLETE, and DISPREFERRED
  - Simple kappa = 0.80
### Tagged Excerpt

<table>
<thead>
<tr>
<th>Time Stamp</th>
<th>Dialogue Stream</th>
<th>Task Stream</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008-04-11 18:23:45</td>
<td>Student: so do i have to manipulate the array this time? [Q]</td>
<td></td>
</tr>
<tr>
<td>2008-04-11 18:23:53</td>
<td>Tutor: this time, we need to do two things [S]</td>
<td></td>
</tr>
<tr>
<td>2008-04-11 18:24:02</td>
<td>Tutor: first, we need to create a new array to hold the changed values [S]</td>
<td></td>
</tr>
</tbody>
</table>

**1-a-i Buggy**

**1-a-i Correct**
Sequence Modeling Task

Given

Qs \rightarrow S_T \rightarrow 1-a-i Corr \rightarrow ?

Predict
Three models

- First-order Markov models (MMs)
- Hidden Markov models (HMMs)
- Hierarchical Hidden Markov models (HHMMs)
Markov model
HMM

Hidden Dialogue State

Dialogue Acts & Task Actions
HHMM

Task/Subtask

Hidden Dialogue State

Dialogue Acts & Task Actions
Portion of Learned HMM

**State 5: Tutor Feedback**
- Stmt (T): 0.05
- Lukewarm Content Fdbk (T): 0.12
- Neg Content Fdbk (T): 0.25
- Positive Fdbk (T): 0.31

**State 1: Student Work**
- Incomplete: 0.15
- Buggy: 0.35
- Correct: 0.38

**State 8: Tutor Hints and Advice**
- Statement(T)-Statement(T): 0.14
- Statement(T): 0.68
Portion of Learned HHMM

Subtask 2 State 4: Tutor Hints and Advice
- PositiveFdbk(T) 0.05
- Question(S) 0.08
- Stmt(T) 0.72

Subtask 2 State 1: Tutor Positive Feedback
- AssessingQuestion(S) 0.09
- PositiveFeedback(T) 0.14
- Incomplete 0.56

Subtask 2 State 6: Student Work And Questions
- Question(S)-Stmt(T) 0.08
- Correct 0.16
- Buggy 0.22
- Incomplete 0.37
Prediction Results by Model

- 1st Order Markov Model: 0.27
- Hidden Markov Model: 0.48
- Hierarchical Hidden Markov Model: 0.57

Baseline: 0.41
HHMM Results by Subtask

<table>
<thead>
<tr>
<th>Task</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understand Problem</td>
<td>0.36</td>
</tr>
<tr>
<td>Subtask 1</td>
<td>0.55</td>
</tr>
<tr>
<td>Subtask 2</td>
<td>0.59</td>
</tr>
<tr>
<td>Subtask 3</td>
<td>0.52</td>
</tr>
<tr>
<td>Subtask 4</td>
<td>0.56</td>
</tr>
<tr>
<td>Extra-domain</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Baseline (.41)
Results in Context

- Catalogue ordering domain
  - Flat model - 55% accuracy (Bangalore et al., 2008)
  - Hierarchical model - 35.6% accuracy (Bangalore & Stent, 2009)

- Our domain, tutoring for introductory computer programming
  - Flat model – 48% accuracy
  - Hierarchical model – 57% accuracy
Conclusions

- Hidden dialogue state useful; HMMs outperform MMs
- Task/subtask structure useful; HHMMs outperform HMMs
- *Understand Problem* phase very challenging to model
- *Extra-domain* conversation most straightforward to predict
Future Work

- Maintain multiple hypotheses for tutorial moves
- Leverage learner characteristics
- Develop fully unsupervised dialogue models
- Ultimate goal: create highly effective data-driven tutorial dialogue systems
A Shameless Plug

The Third Workshop on Question Generation

www.questiongeneration.org/QG2010

June 18, 2010, Pittsburgh
Acknowledgments
Contact: keboyer@ncsu.edu