DISCUSS: Toward a Domain Independent Representation of Dialogue

by

Lee Becker

B.S., Colorado School of Mines, 2000

M.S., University of Colorado, 2010

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2012
This thesis entitled:
DISCUSS: Toward a Domain Independent Representation of Dialogue
written by Lee Becker
has been approved for the Department of Computer Science

Wayne Ward

Martha Palmer

Sarel van Vuuren

James Martin

Tamara Sumner

Barbara Fox

Date ________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
While many studies have demonstrated that conversational tutoring systems have a positive effect on learning, the amount of manual effort required to author, design, and tune dialogue behaviors remains a major barrier to widespread deployment and adoption of these systems. Such dialogue systems must not only understand student speech, but must also endeavor to keep students engaged while scaffolding them through the curriculum. Crafting robust, natural tutoring interactions typically involves writing tightly scripted behaviors for a wide variety of student responses and scenarios.

Combining statistical machine learning with corpus-based methods in natural language processing presents a possible path to reducing this effort. Advances in reinforcement learning have been applied toward dialogue systems to learn optimal behaviors for a given task. However, these learned dialogue policies are tightly coupled to the specific dialogue system implementation. For content-rich applications such as intelligent tutoring systems, there is an immediate need to learn tutoring strategies and dialogue behaviors that can be leveraged across a variety of materials, concepts and lessons. Further generalization will require an intermediate representation of dialogue that can abstract the conversation to its underlying action, function, and content.

This work introduces the Dialogue Schema Unifying Speech and Semantics (DISCUSS), an intermediate linguistic representation that captures the semantics and pragmatics of speech while also allowing for domain-independent modeling of tutorial dialogue. To better understand the benefits of the DISCUSS representation, a corpus of computer-mediated tutorial dialogues was manually tagged with DISCUSS labels. These data were then used for three different tasks: utterance classification, dialogue move selection, and learning gains prediction.

System performance in these tasks demonstrate the utility and viability of the DISCUSS representation for analyzing and automating dialogue interactions. Utterance classifiers achieve
DISCUSS labeling performance on par with inter-annotator agreement levels. System performance in ranking and selecting follow-up questions illustrates the usefulness of DISCUSS-based features for modeling and identifying the factors behind human decision making when teaching. Correlating features of the dialogue with measured learning gains in students shows how DISCUSS-derived metrics provide a detailed account of real tutoring strategies and student behaviors. Together these results represent a step toward more domain-independent mechanisms for modeling dialogue.
Acknowledgements

An endeavor like this would not be possible without the mentorship of an interested thesis advisor; in my case I was lucky enough to have three people overseeing my work. Thank you to my advisors Wayne Ward, Martha Palmer, my PI Sarel van Vuuren, and my committee members Jim Martin, Tamara Sumner, and Barbara Fox for supporting and guiding me over the past several years. Wayne, I wish there was a way to bottle those moments of clarity I had when leaving your office. Your advice always left me supercharged and ready to face any research challenge. Martha, your encouragement and insights always helped me to take a step back and see the larger picture. Sarel, your repeated recommendations to “get empirical” served me well, and I have adopted the expression as a personal research mantra. Jim and Barbara, it has been a pleasure sitting in your classes; your feedback and discussion have played a foundational role in forming this thesis. Tammy, thank you for adopting me as a surrogate member of your own research group.

The hallway conversations and camaraderie at CINC feature prominently in my graduate school memories. Jinho Choi, I could always count on you for a late night chat about NLP and machine learning. Steven Bethard and Philip Ogren, I have learned so much in working on ClearTK with both of you. Shumin Wu, do I need to say more than BBQ Tuesday? Dmitriy Dligach and Phillipp Wetzler, the ML reading group would not have been the same without your eagerness to offer up proofs on the dry-erase board. To Chih How Bong, Ovo Dibie, Jena Hwang, Kirill Kireyev, Daryl Lonon, Praful Mangalath, Keith Maull, Rodney Nielsen, Ifeyinwa Okoye, Manuel Saldivar, and Sudha Verma thank you for all the great discussion on subjects as varied as programming, linguistics, education, psychology, pop culture and food. Again a giant thank you to all members of
the CompSem group and all of my colleagues at CINC.

Mohammad Al-Mutawa, Ashok Basawapatna, Guy Cobb, Holger Dick, Rhonda Hoenigman, Dan Knights, Jeffrey LaMarche, Jane Meyers, and Caleb Phillips, thank you for your friendship. I couldn’t think of a better cohort group with whom to enter CU.

This dissertation would not be possible without the hard work of my linguistic annotators Jennifer Watanabe, Jenette Preciado, and Matthew Goss. Without you, there would be no corpus and thus no thesis.

Festivites, you were there when I first heard about this field of study called computational linguistics, and you encouraged me to pursue a degree in it long before I fully understood what that meant. Thank you for the countless e-mails and for entertaining my numerous musings about language, statistics, and silat.

Kody (Kodiak), it has been reassuring to know that no matter how my day went, I could return home to a wagging tail and soft, fuzzy head that was willing to tolerate a heavy dose of mooshing and hugging.

I would like to thank my mom and brother who were always supportive of my decision to return to school. Joey it has been great having you in Boulder all of these years. Even though we speak different dialects of technical jargon, there’s no geeking out as special as brotherly geeking out. Mom, I am eternally grateful to you for instilling in me a love of learning and a stubborn work ethic. I would not be writing this without all the sacrifices you’ve made for Joey and me. On a lighter note, thank you for putting up with my childish, exasperated responses whenever you asked how my “paperwork” is going. This paperwork is complete Mom.

Most of all, I want to thank my wife Sara MacAlpine for all of her support, encouragement, and love. Sara, I hope that I have been even a tenth as supportive of your own graduate studies as you have been of mine. We have been together for nearly every journey in our adult lives, and I can not imagine traveling down this path without you.

The work presented in this thesis was supported by grants from the National Science Foundation (NSF grant numbers DRL-0733322 and DRL-0733323), the Institute of Educational Science (IES
grant number R3053070434) and the DARPA GALE program (Contract number HR0011-06-C-0022, a supplement for VerbNet attached to the subcontract from the BBN-AGILE Team). Any findings, recommendations, or conclusions are those of the author and do not necessarily represent the views of NSF, IES, or DARPA.
## Contents

### Chapter

1 Introduction
   1.1 Motivation .................................................. 1
   1.2 Thesis Statement ............................................. 4
   1.3 Research Questions .......................................... 4
   1.4 Approach ..................................................... 6
   1.5 Contributions ................................................ 7
   1.6 Organization ................................................ 9

2 Research Context
   2.1 Tutorial Dialogue Setting ................................. 10
      2.1.1 Student Interface ...................................... 11
   2.2 FOSS Curriculum ........................................... 12
   2.3 Dialogue Design ............................................. 13
   2.4 MyST System Architecture ................................. 14
      2.4.1 Semantic Parsing ....................................... 16
      2.4.2 Dialogue Manager ...................................... 17
   2.5 Wizard-of-Oz Study ......................................... 19
   2.6 ASK Assessment Study ...................................... 20
# Prior and Related Work

3.1 Representations of Dialogue

- 3.1.1 Speech Act Theory
- 3.1.2 Conversation Analysis
- 3.1.3 Dialogue Act Taxonomies
- 3.1.4 Tutorial Act and Question Type Taxonomies

3.2 Applications and Techniques

- 3.2.1 Dialogue Acts and Natural Language Processing
- 3.2.2 Tutorial Dialogue Analysis

# DISCUSS: A multidimensional dialogue move taxonomy

4.1 Move Categories

4.2 Semantic Dimensions

- 4.2.1 Semantic Roles
- 4.2.2 Predicate Type

4.3 Pragmatic Dimensions

- 4.3.1 Dialogue Act Dimension
- 4.3.2 Rhetorical Form

4.4 Discussion

# Annotating Tutorial Dialogues with DISCUSS

5.1 Annotators

5.2 Annotation Tool

5.3 Semantic Roles Annotation

5.4 Corpus Statistics

5.5 Inter-Annotator Agreement
6 Automatic Classification of Student Utterances

6.1 Related Work ......................................................... 52

6.2 DISCUSS Classification ........................................... 53

6.3 Selected DISCUSS Labels .......................................... 53
   6.3.1 Merging DISCUSS Labels ................................. 54

6.4 Preprocessing ....................................................... 56

6.5 Features ............................................................. 56

6.6 Model Training, Parameter Selection, and Evaluation .......... 57
   6.6.1 Evaluation ..................................................... 58

6.7 Results .............................................................. 60
   6.7.1 Merged Labels Results ..................................... 63
   6.7.2 Discussion ..................................................... 67

6.8 Error Analysis ...................................................... 68
   6.8.1 Lexical overfitting ......................................... 68
   6.8.2 Context sensitive errors ................................. 69
   6.8.3 Annotator error ............................................. 70
   6.8.4 Evaluation limitations ................................... 71

6.9 Conclusions .......................................................... 72

7 Question Ranking and Selection in Context ............ 74

7.1 Connections to Prior Work ...................................... 75

7.2 Data Collection ..................................................... 76
   7.2.1 MyST Logfiles and Transcripts ......................... 76
   7.2.2 Question Authoring ....................................... 76
   7.2.3 Ratings Collection ....................................... 79
   7.2.4 Rater Agreement .......................................... 80

7.3 Automatic Ranking .................................................. 82
## 9 Conclusions and Future Work

### 9.1 Summary

### 9.2 Research Questions Revisited

### 9.3 Discussion and Future Work

### 9.4 Closing Remarks

## Bibliography

## Appendix

### A DISCUSS Annotation Guidelines

#### A.1 Dialogue Act Definitions

##### A.1.1 Dialogue Control Tags

##### A.1.2 Information Exchange Tags

##### A.1.3 Attention Management Tags

#### A.2 Rhetorical Form Definitions

##### A.2.1 Dialogue Control Tags

##### A.2.2 Information Exchange Tags

##### A.2.3 Attention Management Tags

#### A.3 Predicate Type Definitions

### B DISCUSS Classifiers Error Analysis

#### B.1 Dialogue Acts

##### B.1.1 Acknowledge

##### B.1.2 Answer

##### B.1.3 Close

##### B.1.4 Open

##### B.1.5 Thank
B.1.6 Metastatement ................................................................. 154
B.1.7 SignalNoUnderstanding .................................................... 154
B.1.8 Uninterpretable ............................................................... 154
B.1.9 MergedNoUnderstanding .................................................... 155

B.2 Rhetorical Forms ................................................................. 155
B.2.1 Bye ................................................................. 155
B.2.2 Compare ............................................................. 156
B.2.3 Confirm, YesNo and MergedYesNo ....................................... 157
B.2.4 Define .............................................................. 157
B.2.5 Describe ............................................................ 158
B.2.6 Elaborate ............................................................. 159
B.2.7 Greet ............................................................... 159
B.2.8 Identify .............................................................. 159
B.2.9 Justify .............................................................. 160
B.2.10 List ................................................................. 161
B.2.11 Quantify ............................................................. 162
B.2.12 MergedDescribe ......................................................... 162
B.2.13 MergedYesNo .......................................................... 163

B.3 Predicate Types ................................................................. 164
B.3.1 AcceptRejectMaybe, YesNoMaybe, and MergedYesNoMaybe .................... 164
B.3.2 Activity, Experience, Topic, and MergedActivity ............................... 165
B.3.3 Attribute ........................................................................ 166
B.3.4 CausalRelation, Process and MergedCausalRelation ............................ 167
B.3.5 Configuration ............................................................. 167
B.3.6 Entity ................................................................. 168
B.3.7 Function .............................................................. 169
B.3.8 Location, Route and MergedRoute .......................................... 169
C Correlation with Learning Gains Features

C.1 Basic Features ................................................................. 172
C.2 Phoenix Dialogue Manager Features ...................................... 173
C.3 DISCUSS Features ............................................................. 173
# Tables

Table

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>DISCUSS Dialogue Act Frequencies</td>
<td>45</td>
</tr>
<tr>
<td>5.2</td>
<td>DISCUSS Rhetorical Form Frequencies</td>
<td>46</td>
</tr>
<tr>
<td>5.3</td>
<td>DISCUSS Predicate Type Frequencies</td>
<td>46</td>
</tr>
<tr>
<td>5.4</td>
<td>DISCUSS Inter-Annator Agreement Summary</td>
<td>48</td>
</tr>
<tr>
<td>5.5</td>
<td>DISCUSS Dialogue Act Inter-Annator Agreement</td>
<td>48</td>
</tr>
<tr>
<td>5.6</td>
<td>DISCUSS Rhetorical Form Inter-Annator Agreement</td>
<td>49</td>
</tr>
<tr>
<td>5.7</td>
<td>DISCUSS Predicate Type Inter-Annator Agreement</td>
<td>50</td>
</tr>
<tr>
<td>6.1</td>
<td>Merged DISCUSS Labels</td>
<td>55</td>
</tr>
<tr>
<td>6.2</td>
<td>DISCUSS Classifier Features</td>
<td>57</td>
</tr>
<tr>
<td>6.3</td>
<td>DISCUSS Model Parameters</td>
<td>59</td>
</tr>
<tr>
<td>6.4</td>
<td>DISCUSS Dialogue Act Classifier Performance</td>
<td>60</td>
</tr>
<tr>
<td>6.5</td>
<td>DISCUSS Rhetorical Form Classifier Performance</td>
<td>61</td>
</tr>
<tr>
<td>6.6</td>
<td>DISCUSS Predicate Type Classifier Performance</td>
<td>62</td>
</tr>
<tr>
<td>7.1</td>
<td>Example Dialogue Context</td>
<td>77</td>
</tr>
<tr>
<td>7.2</td>
<td>Inter-Rater Rank Agreement</td>
<td>82</td>
</tr>
<tr>
<td>7.3</td>
<td>Model Features by Category</td>
<td>84</td>
</tr>
<tr>
<td>7.4</td>
<td>Example Learning Goals</td>
<td>86</td>
</tr>
<tr>
<td>7.5</td>
<td>Question Ranking System Scores</td>
<td>88</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>7.6</td>
<td>System Mean Kendall’s-$\tau$ Rank-Order Agreement Scores by Model and Rater</td>
<td>92</td>
</tr>
<tr>
<td>7.7</td>
<td>Distribution of Rater Model Features</td>
<td>93</td>
</tr>
<tr>
<td>7.8</td>
<td>Rater Model Cosine Similarities</td>
<td>93</td>
</tr>
<tr>
<td>8.1</td>
<td>Basic Dialogue Features Correlations</td>
<td>103</td>
</tr>
<tr>
<td>8.2</td>
<td>Phoenix Dialogue Manager Features Correlations</td>
<td>105</td>
</tr>
<tr>
<td>8.3</td>
<td>Sequence Edit Distance Feature Correlations</td>
<td>109</td>
</tr>
<tr>
<td>8.4</td>
<td>Dialogue Act Feature Correlations</td>
<td>111</td>
</tr>
<tr>
<td>8.5</td>
<td>Rhetorical Form Feature Correlations</td>
<td>112</td>
</tr>
<tr>
<td>8.6</td>
<td>Predicate Type Feature Correlations</td>
<td>113</td>
</tr>
<tr>
<td>8.7</td>
<td>DISCUSS Tuple Feature Correlations</td>
<td>114</td>
</tr>
</tbody>
</table>
# Figures

## Figure

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The MyST tutoring interface</td>
<td>12</td>
</tr>
<tr>
<td>2.2</td>
<td>Virtual Human Toolkit architecture</td>
<td>15</td>
</tr>
<tr>
<td>2.3</td>
<td>Example Phoenix Parse</td>
<td>16</td>
</tr>
<tr>
<td>2.4</td>
<td>Example Task Frame</td>
<td>18</td>
</tr>
<tr>
<td>2.5</td>
<td>The MyST WOZ Data Collection Flow</td>
<td>20</td>
</tr>
<tr>
<td>2.6</td>
<td>The MyST WOZ Interface</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>Examples of Speech Acts and Their Effects</td>
<td>24</td>
</tr>
<tr>
<td>3.2</td>
<td>Example Tutorial Dialogue Annotated with DAMSL Acts</td>
<td>26</td>
</tr>
<tr>
<td>4.1</td>
<td>DISCUSS Move Categories Example</td>
<td>35</td>
</tr>
<tr>
<td>4.2</td>
<td>Semantic Role Example</td>
<td>36</td>
</tr>
<tr>
<td>4.3</td>
<td>Predicate Type Example</td>
<td>37</td>
</tr>
<tr>
<td>4.4</td>
<td>Marking Example</td>
<td>38</td>
</tr>
<tr>
<td>4.5</td>
<td>Revoice Example</td>
<td>38</td>
</tr>
<tr>
<td>4.6</td>
<td>Rhetorical Form Example</td>
<td>39</td>
</tr>
<tr>
<td>5.1</td>
<td>Screenshot of DISCUSSed Annotation Tool Propositions Tab</td>
<td>42</td>
</tr>
<tr>
<td>5.2</td>
<td>Screenshot of DISCUSSed Annotation Tool Turns Tab</td>
<td>43</td>
</tr>
<tr>
<td>5.3</td>
<td>Example Frame Element to VerbNet SRL Mapping</td>
<td>44</td>
</tr>
</tbody>
</table>
6.1 *MergedNoUnderstanding* Dialogue Act Precision-Recall Curve . . . . . . . . . . . . . . . . . . 63
6.2 *MergedDescribe* Rhetorical Form Precision-Recall Curve . . . . . . . . . . . . . . . . . . . . . 64
6.3 *MergedYesNo* Rhetorical Form Precision-Recall Curve . . . . . . . . . . . . . . . . . . . . . . 64
6.4 *MergedTopic* Predicate Type Precision-Recall Curve . . . . . . . . . . . . . . . . . . . . . . . 65
6.5 *MergedCausalRelation* Predicate Type Precision-Recall Curve . . . . . . . . . . . . . . . . . . 65
6.6 *MergedRoute* Predicate Type Precision-Recall Curve . . . . . . . . . . . . . . . . . . . . . . . . 66
6.7 *MergedVisual* Predicate Type Precision-Recall Curve . . . . . . . . . . . . . . . . . . . . . . . 66
6.8 *MergedYesNoMaybe* Predicate Type Precision-Recall Curve . . . . . . . . . . . . . . . . . . . 67

7.1 Question Authoring Tool Screenshot . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 79
7.2 Question Rating Form Screenshot . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81
7.3 Distribution of Per-Context Kendall’s-τ Values . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
7.4 Distribution of Per-Context System Ranks . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
7.5 Distribution of Twenty Most Influential Features per Rater Model . . . . . . . . . . . . . . . . . 94

8.1 Example speak/discuss File . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 101
8.2 MyST/Phoenix Task File Snippet . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
8.3 PERCENT_PROMPTS_FROM_RULES Versus Student Learning Gains, Conditioned
   on Pre-test Score. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 107
8.4 PERCENT_STUDENT_TURNS_UNPARSEABLE Versus Student Learning Gains,
   Conditioned on Pre-test Score. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 107
Chapter 1

Introduction

1.1 Motivation

Many studies have shown the benefits of one-on-one tutoring in both human-to-human and computer-to-human contexts (Bloom, 1984; Chi et al., 2001; Cohen et al., 1982; VanLehn et al., 2007). Computer-based, Intelligent Tutoring Systems (ITS) have proven effective for teaching a wide variety of subject matter including programming (Anderson and Reiser, 1985), mathematics, and science. High school and college students taught with an ITS have shown as much as a one-sigma improvement in learning gains (roughly the equivalent of a letter grade) over their peers who learned the same material in a classroom-only setting (Graesser et al., 2001; VanLehn and Graesser, 2002; VanLehn et al., 2005). While the potential benefits of ITSs are understood, the challenge of making this technology effective, accessible and available for a wider population remains. The Computing Research Association and the National Academy of Engineering has even gone as far as identifying personalized tutoring as a grand challenge for the 21st century (“Computing Research Association”, 2003; “National Academy of Engineering”, 2008)

Intelligent tutoring systems are becoming increasingly as effective as one-on-one human tutors (VanLehn, 2011). Advances in speech and language technology have a played a role in this improvement by providing “finer user interface granularity” (VanLehn, 2011). Improved speech recognition and natural language understanding have given rise to a new generation of natural language ITSs (Graesser et al., 2001; VanLehn and Graesser, 2002; VanLehn et al., 2005; Ward et al., 2011b) that have moved beyond the keyboard towards immersive multi-modal environments.
that integrate language with audio and visual interfaces. These ITSs facilitate learning by engaging learners through conversation with a virtual tutor and through interaction with multimedia content.

Despite these advances, ITSs are still limited by their ability to conduct robust, natural conversations. Improving natural language dialogue is an active area of research in both ITS and computational linguistics. A core problem in spoken dialogue system research is the task of dialogue management, that is, the selecting of the system’s next dialogue move or action. In most ITSs creating dialogue management strategies is still a manual, labor-intensive process. Creating dialogue behavior robust enough to handle a wide variety of responses still requires many man-hours of manual authoring and fine tuning of the dialogue actions before a system is ready for deployment. In many cases the authors may need programming knowledge to implement the desired behavior. Authoring tools like TuTalk (Jordan et al., 2007) and the AutoTutor Authoring Tool (Susarla et al., 2003) help to streamline the dialogue authoring process, but this process is still restricted by the underlying dialogue models. AutoTutor’s dialogues strictly adhere to a hint, prompt, assert sequence while TuTalk’s approach is built upon a finite state machine (FSM) methodology. Though these techniques work well when the prompts are clear and direct, they can be constraining when the tutorial style demands a more open-ended line of questioning. Even after authoring a dialogue, it is difficult to know which tutor actions and questions play a role in making the dialogue more successful, and experimenting with different pedagogical styles requires a complete re-authoring.

To fully capitalize on the potential benefit of ITS-based learning will require moving beyond authoring tightly-scripted behaviors to developing capabilities for handling richer, more natural, open-ended tutorial dialogues. Machine learning techniques for dialogue management have received increased attention with advances in probabilistic reasoning such as reinforcement learning (Singh et al., 2002) and Markov decision processes (Williams and Young, 2005). Unlike hand-crafted approaches, which often require exhaustive testing for coverage of all possible behaviors, these approaches learn behaviors directly from human dialogue corpora. Recently, these techniques have been successfully applied to data from tutorial dialogue systems (Chi et al., 2009, 2010; Tetreault and Litman, 2008). However these successes have yet to be widely deployed, as the
behavior is only optimized over a small number of dialogue decisions and is still not tractable for more open-ended dialogues. Furthermore, the learned dialogue strategies are tightly coupled to a specific system implementation. Advancements in tutorial dialogue management will not only require improvements in the machine learning algorithms, but will also require an information model that supports a more general dialogue state. Such an information model will require an intermediate representation of dialogue that abstracts it to its underlying action, function, and content.

Parallel to the goals of improving dialogue management, there is also a desire to understand what characteristics of the dialogue make for successful tutoring sessions. As stated by Litman and Forbes-Riley (2006):

Research in tutorial dialogue systems is founded on the belief that a one-on-one natural language conversation with a tutor provides students with an environment that exhibits characteristics associated with learning. However, it is not yet well understood exactly how specific student and tutor dialogue behaviors correlate with learning, and whether this generalizes across different types of tutoring situations.

Though there is a growing body of work wherein dialogue behavior and dialogue representations are correlated with learning (Graesser and Person, 1994; Litman and Forbes-Riley, 2006; Litman et al., 2009), most of these analyses use coarse tutorial acts and do not give insight into the content, form or function of the tutor’s questions and student’s responses. As with the desire to learn generalized strategies, the goal of dialogue characterization and analysis hinges on having an expressive, linguistically-motivated dialogue representation. From an analysis perspective, such a representation would provide a more detailed account of conversational transactions, while from a generative point of view, it could provide more systematic control for comparing and contrasting differences in pedagogical styles.
1.2 Thesis Statement

Advances in both dialogue management and intelligent tutoring systems will require a new representation of dialogue to bridge between data-driven techniques and rules-based approaches. Such a representation should be descriptive enough to facilitate analysis that yields insight into the nature of tutoring and task-oriented dialogue. At the same time, it should be abstract enough to allow for generalization across multiple domains. Existing dialogue representations only capture coarse-grained actions, while sentence-level semantic representations are too specific to capture the relation between speaker’s turns. A complete depiction of dialogue interaction would span this gap with dimensions that merge dialogue action and intent with additional syntactico-semantic layers. This thesis argues that the Dialogue Schema Unifying Speech and Semantics (DISCUSS) provides a powerful representation for analyzing and automating interactions within tutorial dialogues. By accounting for speakers’ action, function, and content, DISCUSS provides a richer account of tutoring than existing dialogue act and tutorial act taxonomies.

1.3 Research Questions

This thesis contributes to the overarching goals of 1) enabling data-driven creation of dialogue systems and 2) creating a research framework suitable for analyzing, discovering and comparing differences in tutoring and pedagogical strategies. Toward that end, this thesis focuses on developing a linguistically-motivated dialogue move taxonomy for annotation of utterances in tutorial dialogues and demonstrating the utility and feasibility of this representation through empirical evaluation of computational models constructed from a corpus of computer-mediated tutorial dialogues. These goals give rise to the research questions listed below.

Research Question 1. What semantic and pragmatic representations are necessary for modeling tutorial dialogue and at what granularity?

Hypothesis 1.1 Dialogue acts alone are too coarse to capture the types of prompts and
responses seen in tutoring. Conversely, sentence-level semantic representations are too specific for modeling conversational discourse. Additional syntactico-semantic information is needed to bridge between this gap and to imbue dialogue moves with a more complete account of the action, function, and content contained within a dialogue. (Chapters 3 and 4)

**Hypothesis 1.2** Learning domain-independent dialogue behaviors from data will require a way to generalize the task-specific semantics. (Chapters 4, 7 and 8)

**Research Question 2.** Are DISCUSS categories well enough defined to attain usable inter-annotator agreement?

**Hypothesis 2.1** Based on existing literature on dialogue move annotation, linguists (with sufficient training) should show high inter-annotator agreement on dialogue act annotation. While annotating discourse moves and semantic predicate categories presents more opportunities for confusion, annotators should achieve fair but usable inter-annotator agreement. (Chapter 5)

**Hypothesis 2.2** Computational models trained to label dialogues with a rich dialogue move taxonomy should achieve results in line with inter-annotator agreement scores. (Chapter 6)

**Research Question 3** How do DISCUSS-based features aid in dialogue decision making tasks such as selecting follow-up questions?

**Hypothesis 3.1** Questioning styles and preferences vary from tutor to tutor, even when teaching with the same pedagogical philosophy. Modeling these individual preferences requires a range of features from shallow lexical features to more complex semantic, pragmatic, and dialogue context features. (Chapter 7)

**Research Question 4** What additional insights can a rich, dialogue move representation like DISCUSS provide for dialogue analysis tasks such as predicting learning gains from dialogue transcripts?
Hypothesis 4.1 While shallow measures such as student utterance length have been shown to correlate with learning, exploring more detailed phenomena like the role of self-explanation in learning requires more detailed information about the kinds of questions tutors ask and the answers students provide. (Chapter 8)

1.4 Approach

Supervised machine learning is employed throughout this thesis as a framework for investigating the properties of the DISCUSS representation. This methodology is commonly used in corpus-based natural language processing and computational linguistics research, as it provides an empirically-grounded approach to testing hypotheses about the data. At a conceptual level, supervised machine learning is the task of inferring a function from labeled training data. Within speech and language processing the learned function typically serves one of three purposes: 1) to label raw text with categories, 2) to overlay linguistic structure over unstructured information or 3) to provide an actionable decision.

A typical breakdown of the stages in this approach consists of data collection, data annotation, feature extraction, model training and evaluation. For the work presented in this thesis, much of the effort centers on manually annotating a corpus of tutorial dialogue transcripts with DISCUSS labels (data collection and annotation stages). These data are then used either as gold-standard labels during training (output) or as an intermediate representation that can be manipulated and consumed for feature extraction (input). Features obtained during the feature extraction stage serve as hypotheses about the relative impact of variables in producing the desired impact for some task. Because this work comes from a natural language processing tradition, there is an emphasis on designing and developing features that can be automatically extracted from the text and/or DISCUSS representations.

This dissertation is comprised of three main tasks: 1) student utterance classification, 2) question ranking and selection, and 3) correlating features of the dialogue with learning gains. Task 1 serves to demonstrate that DISCUSS labeling can be done automatically, while tasks 2 and 3 were
selected to substantiate the usefulness of the representation for dialogue automation and analysis.

The student utterance classification task uses the text of a student’s turn as input, and labels the turn (utterance) with the appropriate DISCUSS tags. Data collection for this task began with a corpus of computer-mediated tutorial dialogues collected from the MyST intelligent tutoring system. These dialogues were then manually annotated with DISCUSS tags. The utterances’ DISCUSS tags then serve as the gold-standard output for training classifiers for automatic DISCUSS-tagging of dialogues. These gold standard data are paired with features extracted from the utterances themselves as well as their associated dialogue context to learn classifier models for each DISCUSS label.

In the question ranking and selection task the goal is to select a follow-up question from a pool of predefined questions for a given dialogue context. To create gold standard data, experienced human tutors rated the questions on a 1-10 Likert scale. These scores were used to create rank-orderings, which are then paired with features extracted from the questions and corresponding dialogue context. These data are then used to learn a preference function for question ranking / selection. This task was selected to illustrate how DISCUSS-based features assist with dialogue decision making tasks, and it serves as a platform for analyzing what factors and cues drive decision making during tutoring.

While the question ranking task shows DISCUSS’s utility for localized decision making, the correlation with learning gains task is a framework intended to highlight the insights the DISCUSS representation affords for higher-level analysis and characterization of dialogues and dialogue quality. Because the tutoring sessions used for this study were not part of the pool of DISCUSS-annotated dialogues, the models for student utterance classification were run to obtain DISCUSS labels for feature extraction. For this work, features and statistics extracted from dialogue transcripts are correlated with measured learning gains obtained from pre- and post-tutoring assessment exams.

### 1.5 Contributions

The major contributions of this work are as follows:
• **The DISCUSS representation:** The Dialogue Schema Unifying Speech and Semantics (DISCUSS) is a linguistically-motivated dialogue move representation that synthesizes lessons from existing dialogue move, tutorial act, and question taxonomies into a rich, multidimensional annotation scheme that captures the underlying action, function, and content contained within a dialogue. This representation provides useful information for making localized decisions within the course of a dialogue as well as for global-level analysis of dialogues. Additionally, this can be used as an intermediate representation for question generation.

• **DISCUSS Annotated Corpus of Tutorial Dialogues:** In total 121 tutorial dialogue transcripts each of approximately 15 minutes in length were manually annotated with the DISCUSS representation. Inter-annotator agreement statistics indicate sufficient reliability and give confidence that a computer can conduct the same labeling task. As a resource, this corpus can support additional queries into the language and techniques used in tutoring.

• **Computational Model for Automatic Tagging of Student Utterances with DISCUSS labels:** This work demonstrates the practicality of the DISCUSS representation, and the individual classifiers enable DISCUSS analysis on unlabeled dialogues. Error analysis helped to identify issues with the DISCUSS representation and resulted in refined categories that produced higher labeling accuracy.

• **Statistical Model for Ranking Follow-up Questions:** This model illustrates the utility of the DISCUSS representation for making localized dialogue decisions, and shows how pairwise ranking and supervised machine learning can be used to select dialogue moves. Analysis of the results gives insight into the factors behind different tutors’ preferences and styles in questioning. Additionally, this work demonstrated a new way to combine human preference/judgment data with Wizard-of-Oz data to learn dialogue decision-making behavior.
• **Correlation-based Analysis of Learning Gains and Dialogue** This work demonstrates the benefit of DISCUSS-derived features for global-level characterization of dialogue. Computed correlations confirm existing notions about tutoring and closer analysis of these metrics uncovers artifacts of the design of the MyST tutoring system. Most importantly, these results inform future experimental designs for comparing and contrasting tutoring styles in dialogue.

1.6 **Organization**

The remainder of this document is structured as follows. Chapter 2 provides background on the MyST intelligent tutoring system and describes the data used throughout this dissertation. Chapter 3 presents existing work in dialogue modeling, dialogue act representations, and tutorial act taxonomies. Chapter 4 introduces the reader to the DISCUSS dialogue move representation. Chapter 5 describes the data collection and annotation studies that produced the corpus of MyST transcripts annotated with DISCUSS tags. Chapter 6 presents the computational models for automatically labeling student utterances with DISCUSS. Chapter 7 introduces the reader to the task of ranking questions in context and describes the machine-learned models trained for the task. Chapter 8 presents analyses of correlations between dialogue features and measured learning gains. Chapter 9 revisits the thesis statement and research questions, and discusses directions for future work.
Chapter 2

Research Context

The research, data collection, and methods in this dissertation were all carried out within the context of an intelligent tutoring system dubbed My Science Tutor (MyST) (Ward et al., 2011b). This system is a conversational virtual tutor designed to improve science learning and understanding for students in third, fourth, and fifth grades. The remainder of this chapter describes the MyST project, its underlying pedagogical philosophies, and the data collected from year-long assessment studies. It is intended to aid the reader in understanding the setting and motivations that led to the design of the DISCUSS dialogue representation.

2.1 Tutorial Dialogue Setting

The goal of the MyST project was to help struggling students learn the science concepts encountered in classroom science instruction through the use of tutorial dialogues. Students using MyST investigate and discuss science through natural spoken dialogues and multimedia interactions with a virtual tutor named Marnie. The MyST dialogue design and tutoring style is based on a pedagogical approach called Questioning the Author (QtA) (Beck et al., 1996; Beck and McKeown, 2006). QtA is an approach to teaching reading comprehension based on the simple premise of encouraging students to grapple with and reflect upon what the author is trying to express. This mode of thinking is designed to force students to form their own understanding of the material. Teachers facilitate this discovery by challenging students with open-ended questions and by directly

1 Parts of this chapter were adapted from My Science Tutor a Conversational Multimedia Virtual Tutor for Elementary School Science, ACM Transactions on Speech and Language Processing (Ward et al., 2011b)
keying in on ideas expressed in the students’ language. A common QtA prompt may take the form “It sounds like you’re talking about X. Can you tell me more about that?” In the context of an inquiry-based science curriculum, “Questioning the Author” becomes “Questioning the Observations and Data”, and each MyST dialogue centers around a set of key concepts that the student is expected to know.

Each 15 to 20 minute MyST dialogue session functions as an independent learning activity that provides, to the extent possible, the scaffolding required to stimulate students to think, reason and talk about science during spoken dialogues with the virtual tutor Marnie. The goal of these multimedia dialogues is to help students construct and generate explanations that express their ideas. The dialogues are designed so that over the course of the conversation with Marnie, the student is able to reflect on his explanations and refine his ideas in relation to the media he is viewing or interacting with, which ultimately leads to a deeper understanding of the science being discussed.

The subjects for MyST’s data collection and assessment come from a pool of students from multiple schools in the Boulder Valley School District. Student participation with MyST was purely optional and carried out independently of class. MyST tutoring is not used as a replacement for class time, and is used as a supplement after they have covered the subject matter in a traditional classroom setting.

2.1.1 Student Interface

Conversations with Marnie are characterized by two key features: the inclusion of media, in the form of an illustration, animation or interactive simulation throughout the dialogue, and the use of open-ended questions related to the phenomena and concepts presented via the media. Students interact with the MyST system via spoken language and point and click input. Wearing microphone equipped headphones, the students engage in conversation with the Marnie avatar who acts as their tutor for the lesson. To limit the influence of literacy rates on the tutoring experience, dialogue is presented purely through audio. Early versions of MyST used for Wizard of Oz data collection
used a text-to-speech synthesizer, while the standalone MyST system used in the assessment study used pre-recorded audio. The bulk of MyST screen real estate is dedicated to the presentation of interactive multimedia. An example screenshot of this interface is shown in Figure 2.1.

![Image of the MyST tutoring interface](image)

**Figure 2.1: The MyST tutoring interface**

### 2.2 FOSS Curriculum

The material presented in MyST follows a curriculum known as the Full Option Science System (FOSS). FOSS is an inquiry-based science program that is based on the idea that “The best way for students to appreciate the scientific enterprise, learn important scientific concepts, and develop the ability to think well is to actively construct ideas through their own inquiries, investigations and analyses.” FOSS is in use in every state in the United States by over 100,000 teachers and 2 million students, and it has been under development since 1988 by the Lawrence Hall of Science, University of California Berkeley.

The FOSS grades 3-6 curriculum consists of twenty\(^2\) modules covering topics as diverse as

\(^2\) [http://www.lhsfoss.org/scope/index.html](http://www.lhsfoss.org/scope/index.html)
Water, The Human Body, and Solar Energy. Each module consists of 4 investigations, and each investigation typically has 3-4 parts. Individual MyST tutorial dialogues correspond to a single part, with the learning goals and many of the multimedia visuals drawn directly from the experiments the students run in class. While FOSS is a complete K-8 curriculum, MyST currently targets only a subset of the modules (Water, Magnetism and Electricity, Measurement, Variables) for grades 3-5.

The learning objectives in each FOSS module are aligned to the National Science Education Standards and standards for most states. The instructional materials for each module are packaged in a kit that contains the materials needed to conduct classroom science investigations, a teacher guide, a module-specific teacher-preparation video, a summative assessment (ASK: Assessing Science Knowledge) to be administered before and after each module, and a set of stories related to the module’s science content. In addition, the FOSS web site, FOSSweb (FOSS), provides resources related to each science investigation including training videos, learning objectives, multimedia activities, glossaries and home activities.

Within a science module, students in classrooms work in small groups to conduct a series of 4 to 5 science investigations over an 8 to 10 week period. Each science investigation consists of a sequence of investigation parts; these are the individual hands-on classroom investigations aligned to specific science concepts and learning objectives. The FOSS program provides an ideal test bed for research and evaluation of MyST, as MyST dialogs are aligned with classroom science investigations, which are aligned to specific learning objects, science standards, and ASK assessments.

2.3 Dialogue Design

Each tutorial session in MyST is designed to cover a few main points (approximately 2 to 4) in a 15 to 20 minute session with a student. The tutorial dialogue is designed to get students to articulate concepts and be able to explain processes underlying their thinking. Tutor actions are designed to encourage students to share what they know and help them express why they know what they know. For the system (Marnie), the goal of a tutorial session is to elicit responses from students that show their understanding of a specific set of points. More specifically MyST
dialogue is a sequence of mini-conversations through a set of semantic frames. Each semantic frame corresponds to a concept or factual proposition, and MyST dialogue aims to elicit student responses that entail these propositions. The approach to eliciting these responses is heavily influenced by QtA techniques and strategies with a heavy emphasis on using prompts which promote self-expression and encourage students to think more deeply about concepts. Two commonly used QtA strategies employed by MyST are marking and revoicing. These two techniques require the ability to identify the student’s dialogue content (referred to as marking it) followed by repeating (revoicing) the question back to the student using similar phrasing; for instance, “You mentioned that electricity flows in a closed path. What else can you tell me about how electricity flows?”

The interactions within a frame typically begin with open-ended questions about the target concept. Further sequences are written in such a way that they proceed from more general open-ended questions (“What’s this all about?”) to more directed open-ended questions (“Tell me more about the flow of electricity in the circuit”). Initially, students are prompted to consider a concept in terms of their recent experiences in class. More details about the specific mechanisms enabling this behavior can be found below in Sections 2.4.1 and 2.4.2.

### 2.4 MyST System Architecture

MyST was developed using Boulder Language Technologies’ (BLT) Virtual Human Toolkit (VHT). The BLT VHT is a resource for designing and experimenting with multimedia programs that support real time conversational interaction with virtual humans. The VHT provides a general purpose platform, a set of technology modules, and tools for researching and developing conversational systems using natural mixed initiative interaction with users in specific task domains. In mixed-initiative dialogs, either the user or the system can seize the initiative and take control of the dialog. The toolkit consists of an integrated set of authoring tools and technologies for developing applications that incorporate virtual humans in applications. It provides authoring tools for presenting and interacting with media (text, images, audio, video and animations), designing and controlling lifelike 3-D computer characters (Cole et al., 2003), and designing natural spoken
dialogs with the virtual agent. A diagram of the VHT architecture is listed in Figure 2.2, and the VHT is composed of the following modules:

- speech recognition
- speech synthesis
- semantic parsing
- dialogue management
- character animation

![Virtual Human Toolkit architecture](image)

**Figure 2.2: Virtual Human Toolkit architecture**
2.4.1 Semantic Parsing

The Phoenix parser (Ward, 1994) maps the speech recognizer output, or any text, onto a sequence of semantic frames. These frames represent the system’s understanding of an utterance. The type of representation Phoenix uses to extract information from user input is generally referred to as shallow semantics. Shallow semantics represents the entities, events and relations between them important to understanding an utterance. In Phoenix, these are characterized as semantic frames, together with semantic frame elements. An example parse for the utterance *Electricity goes from minus to plus* appears in Figure 2.3

<table>
<thead>
<tr>
<th>Frame: FlowDirection</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="electricity">Electricity</a></td>
</tr>
<tr>
<td><a href="goes">Flows</a></td>
</tr>
<tr>
<td>[DirFlow].[Origin].Negative(minus)</td>
</tr>
<tr>
<td>[DirFlow].[Destination].Positive(plus)</td>
</tr>
</tbody>
</table>

Figure 2.3: Example Phoenix Parse

Semantic grammars are used to match word strings against patterns for frame elements. These are context free patterns where the non-terminal nodes are concepts, events and relations important in the domain. Separate grammars are written for each frame element (like [DirFlow].[Origin]). In matching frame element grammar patterns against the input text, the parser ignores words that do not match any frame element. This allows the system to match expressions relevant to understanding the domain while ignoring extraneous information and disfluencies such as restarts. A Viterbi search is used to find the optimal set of frames and frame elements. The most optimal parse is the one that covers most of the input and is least fragmented. A set of parses of equal score is produced for an ambiguous input. The grammar rules may be written manually or may be trained from an annotated corpus if one is available.
2.4.2 Dialogue Manager

The Dialogue Manager (DM) controls the system’s dialogue interaction with the user and is responsible for: (a) maintaining a context representing the history of the dialog; (b) selecting a preferred parse from a set of candidate parses given the context; (c) integrating the new parsed input into the context; (d) generating a sequence of actions based on the context. The DM also uses the frame representation used by the parser.

MyST’s dialogue behavior, including the presentation of media within dialogues, is controlled by a task file (Pellom et al., 2000). The task file contains the definition of the task frames to be used by the application. A task frame is a data object that contains all of the information necessary (or at least available) to interact about the frame. The building blocks for a task file include:

- Frame Elements – the extracted information, also known as slots.
- Templates for generating responses
- Pattern-Action pairs, called rules, for generating responses contingent on certain conditions in the context.

By default, MyST will attempt to elicit speech to fill the frame elements representing the propositions of a frame. A sequence of interface actions is generated to elicit a response. The set of interface actions used are: flash(), movie(), show(), clear(), speak() and synth(). An example action sequence would be flash(Components); synth(Tell me about that.). This sequence would run the Flash file Components and would synthesize the word sequence and have the avatar (Marnie) speak it. In order to elicit speech to fill a frame element, the system developer specifies a list of action sequences for the element. During a session, the Dialogue Manager (DM) keeps count of how many times each element has been prompted for and uses the next action sequence in the list. Once it has exhausted the list, it gives the element the value FAIL, and will move on. Figure 2.4 shows an example of a task frame.

The tutorial dialogue author may also specify a set of rules for the frame. Rules are pattern-
action pairs that can be used to generate action sequences conditioned on features of the context. Rule pattern definitions are boolean expressions based on element values in the context. If the rule evaluates to true, one of the action sequences following it are sent to the interface manager. Like when prompting for an element, the system keeps count of the number of times a rule has been used and uses the next sequence each time. The rule shown in Figure 2.4 will trigger its associated prompts when the student reverses the origin and destination slots within the Flow frame. Within MyST, rules are the primary mechanism for enabling QtA revoicing and marking behavior.

```
Frame: FlowDirection
Description: Electricity flows from the negative terminal to the positive terminal of a battery.
[Flow]
[DirFlow]
  Action: flash(Flow); synth(Tell me about what’s going on here.)
[DirFlow].[Origin]
  Action: synth(What do you notice about the flow?)
[DirFlow].[Destination]
  Action: Action: flash(Flow); synth(Which side of the battery is the electricity going to)

Rules:
# Student got direction backward
([DirFlow].[Origin] == "positive") || ([DirFlow].[Destination] == "negative")
  Action: flash(Flow); synth(Tell me again about the flow?)
  Action: flash(Flow); synth(What direction is it going?)
```

Figure 2.4: Example Task Frame

The DM uses a stack driven algorithm for flow control. It maintains two frame stacks, 1) current, the set of currently active frames, and 2) history, the set of completed frames. The DM tries to complete the frame on top of the current stack. If the frame on top is complete, it is moved to the history stack and the new top frame is completed. In attempting to complete a frame, the rules are checked first. If a rule expression evaluates TRUE and it has not been marked FAIL, the next action sequence for the rule is used. If no sequence was generated by checking the rules, the DM determines the first unfilled frame element that has an associated action sequence. If all required elements are filled, the frame is moved to the history stack, and the system attempts to fill the new top frame. The action sequences for both rules and frame elements can cause new frames to be pushed onto the current stack, or old frames to be moved off to the history stack.
2.5 Wizard-of-Oz Study

To gather data for MyST system coverage and dialogue analysis, researchers at Boulder Language Technologies conducted Wizard-of-Oz (WOZ) experiments that allowed a human tutor to be inserted into the interaction loop. Sessions were monitored by project tutors (former science teachers trained in QtA-style tutoring) who served as Wizards. Wizards were responsible for accepting and overriding any system action as well as tracking which target proposition (learning goal) was currently in focus. The wizards in this study are a pool of paid tutors with former teaching experience. Prior to leading WOZ sessions, the tutors were trained on best practices for both Questioning the Author and the FOSS Science Curriculum.

Over the past three years the MyST project has accumulated over five-hundred, 15-minute WOZ sessions spread across four modules Magnetism and Electricity, Measurement, Variables, and Water each with 16 lessons. In total 392 students participated in MyST tutoring. Student speech from these sessions was professionally transcribed at the word level. Disfluencies (false starts, truncated words, filled pauses, etc.) are also marked in the transcriptions. Each WOZ dialogue session produces a log file with time-stamped entries for the events that occurred during the dialogue. After the speech has been transcribed, the transcription is merged with the log file to give a full account of the dialogue session. Figure 2.5 illustrates this flow and Figure 2.6 shows an example of the interface used by the wizards.

Each MyST dialogue session produces a log file that contains time-stamped entries for the events that occurred during the dialog. At each point that the student speaks, an entry is written into the log that gives the filename for the associated recorded speech file. The speech recognition output is logged. Manual transcription of the speech files is performed offline and is introduced into the log file later. Some additional pieces of information stored in the log file are: extracted frame elements, current context, frame name, and frame element or rule that is generating the system response, the number of times this frame element or rule has been used, and the action sequence generated for the response.
2.6 ASK Assessment Study

Over the 2010-2011 school year, researchers on the MyST project carried out a long-term study to assess the effect of their intelligent tutoring system (Ward et al., 2011b). The experimental
design compared students receiving computer-assisted tutoring with those receiving face-to-face human tutoring. The transcripts and data analyzed in this chapter come exclusively from the virtual tutoring condition.

Students in the study received in-class instruction in the FOSS modules and then participated in a series of 16 conversations with MyST lasting 15 to 20 minutes per session. Together these sessions cover the material for one FOSS module. To ensure MyST sessions were reviewing rather than teaching new material, MyST deployment occurred after a classroom had completed a specific FOSS investigation. For example, subjects would not converse with MyST about parallel circuits until the equivalent lecture and laboratory had been completed in the classroom. In total one-hundred-two (102) students in 14 classrooms in 6 schools participated in MyST tutoring (standalone condition) for the assessment study.

To measure learning gains, the ASK (Assessing Science Knowledge) test was administered to students before (pre-test) and after (post-test) each module. ASK is a standard part of the FOSS curriculum and is bundled as part of the classroom kits given to teachers. The ASK assessments for the four modules used in the assessment have identical pre and post versions. Questions included open-ended, short answer, multiple choice and graphing items. Students took tests before the beginning of FOSS lessons, and immediately after tutoring ended at the school. Pairs of raters (tutors who had previously served as Wizards in the WOZ study), scored all assessment exams following scoring rubrics provided by FOSS. Inter-rater reliability was high with intra-class correlation coefficients ranging from 0.89 to 0.98. Internal reliabilities using Cronbach’s Alpha were lower with $\alpha_{pre} = 0.74$ and $\alpha_{post} = 0.79$. The final scores used for outcome analysis were the averages across both raters. While these data were collected primarily to validate the efficacy of a conversational intelligent tutoring system, they also form the basis for this thesis’ exploration of what specific factors of tutor and student dialogue behavior correlate with learning gains.
Chapter 3

Prior and Related Work

Abstracting dialogue into analyzable, functional units has long been a goal within the fields of linguistics, artificial intelligence and natural language processing. Similarly research in education, the learning sciences, and intelligent tutoring systems have long sought dimensions for categorizing questions, learning goals, tutorial strategies and student behavior. The dialogue models described in this thesis center on a new dialogue move taxonomy known as the Dialogue Schema Unifying Speech and Semantics (DISCUSS). This taxonomy is an attempt to combine the above notions about tutoring and dialogue into a framework suitable for enabling analysis and automation of tutorial dialogue.

Resolving the distinct, but overlapping goals of dialogue modeling and tutorial analysis requires understanding the issues important to the fields of computer science, linguistics and education. Creating DISCUSS was a process of managing competing aims of linguistic richness, analytical effectiveness, dialogue generation capability, and discriminative power for machine learning. This process was heavily influenced by the existing literature in dialogue representations. The remainder of this chapter highlights prior work covering the current state of dialogue act modeling and tutorial dialogue analysis. Together this body of research helps to define the desired depth, breadth, and organization of the DISCUSS dialogue move representation.
3.1 Representations of Dialogue

3.1.1 Speech Act Theory

Speech Act Theory (Austin, 1962; Searle, 1969) is one of the earliest works to propose an intermediate representation of dialogue. Austin (1962) coined the term *speech act* to describe how a speaker’s words and utterances performs action, and subsequently identified three classes of speech act: *locutionary* acts, *illocutionary* acts, and *perlocutionary* acts. Locutionary acts capture the utterance’s surface meaning or literal content, illocutionary acts describe the semantic force or intention of the utterance, and perlocutionary acts indicate the utterance’s effect. Within a single utterance it is possible to simultaneously extract interpretations from each of these modes. This division allows for separation between the speaker’s intent and the propositional content contained within his speech. Thus, for an utterance like “Did you take out the trash?”, the speech act account yields two levels of interpretation. At a locutionary act level, this is simply as a yes or no question. From a illocutionary act perspective, this utterance serves as a command or suggestion to the listener to take action. Searle (1975) further refined this notion by placing speech acts into five basic categories:

- **Representatives** speech acts that commit a speaker to the truth of the expressed proposition (e.g. assertions, conclusions, etc.)

- **Directives** speech acts that cause the hearer to take a particular action, (e.g. requests, commands and questions)

- **Commissives** speech acts that commit a speaker to some future action, (e.g. promises and oaths)

- **Expressives** speech acts that express the speaker’s attitudes and emotions towards the proposition (e.g. congratulations, thanks, and apologies)

- **Declarations** speech acts that change the institutional state of affairs and often rely on extra-linguistic contexts (e.g. christenings, declarations of war, firing from employment)
Allen, Cohen, and Perrault brought Speech Act Theory into the realm of artificial intelligence and planning by defining preconditions and effects relating to execution of different speech acts Allen and Perrault (1980); Cohen and Perrault (1979). An example of these effects are shown in table 3.1.

| REQUEST(speaker, hearer, act): | effect: speaker WANT hearer DO act |
| INFORM(speaker, hearer, proposition): | KNOW(hearer, proposition) |

Figure 3.1: Examples of Speech Acts and Their Effects

When using the original set of speech acts for dialogue modeling, Allen and Perrault (1980); Cohen and Perrault (1979); Bratman et al. (1988) found the model’s expressiveness and functionality were restricted by the coarse granularity of the speech acts themselves. Additionally, the effects used by their planning systems needed to be manually authored, and were not necessarily representative of real conversation.

### 3.1.2 Conversation Analysis

In contrast to speech act theory, conversation analysis (CA) (Sacks et al., 1974; Schegloff, 2007) is neither concerned with words as semantic or discourse units nor focused on speaker intent. Instead CA aims to describe and discover the emergent properties and structures that arise from natural language interactions. In particular CA focuses on dialogue phenomena including:

- turn-taking (Sacks et al., 1974) - the process by which speakers allocate or defer to participate in conversation
- adjacency pairs (Sacks et al., 1974) - the natural organization of speech into responsive pairs; though pairs may be split across several terms
- repair (Schegloff et al., 1977) - how parties in conversation deal with problems in speaking, hearing, or understanding
• preference organization (Pomerantz, 1984) - the ways through which different types of conversational actions are carried out sequentially. For example whether a response to one utterance is 'preferred' vs. 'dispreferred'.

From a computational perspective, CA has been mainly limited to the task of detecting and responding to moments for dialogue repair (Cawsey, 1991; McTear, 1985; Yang, 2005). Though DISCUSS takes a more speech-acts oriented approach to dialogue modeling, many aspects of DISCUSS design were motivated by a desire for a concise way to spot dispreferred responses in adjacency pairs.

3.1.3 Dialogue Act Taxonomies

While speech act theory forms the foundation for many models of dialogue, it was not created with computational tasks in mind. The decades following Austin and Searle’s original works have brought several refinements to speech act theory in the form of various dialogue act taxonomies.

3.1.3.1 DAMSL

Core and Allen created the DAMSL (Dialog Act Markup in Several Layers) (1997) taxonomy to facilitate automatic analysis of dialogues and computer participation in dialogues. The designers of DAMSL argued that a major drawback to speech act theory is the limitation it imposes on the number of acts per utterance. Their claim focused on how an act can perform multiple acts simultaneously. To address this shortcoming, DAMSL was designed with three layers: Forward Communicative Functions, Backward Communicative Functions, and Utterance Features. The Forward Communicative Functions closely resemble Searle’s taxonomy with categories for Representatives, Directives, and Commisives. The Backward Communicative Functions exist separately from the Forward dimension and are used for showing that an utterance relates to another utterance. Backward acts are typically representative of responses that show agreement, understanding or that convey an answer. Lastly, DAMSL’s Utterance Feature layer simultaneously captures aspects of the underlying semantic content while also hinting at the surface form of the utterances.
While DAMSL’s acts are general enough to port across different domains, they are not descriptive enough to support natural language generation of utterances. Furthermore, for complex task-oriented domains like tutoring, DAMSL provides little insight into how information is exchanged. Consider the example tutorial dialogue in figure 3.2. The DAMSL tags do not convey much about the exchange save for that it is a series of questions and answers. To answer questions about whether or not a tutoring session is successful or whether the tutor is attending to the needs of the student will require a richer representation. Such a representation should be able to differentiate questions and answers by their function and should link pragmatic intentions to semantic structures.

<table>
<thead>
<tr>
<th>1G Turn</th>
<th>Speaker</th>
<th>Utterance</th>
<th>Forward</th>
<th>Backward</th>
<th>Utterance Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tutor</td>
<td>Tell me about magnetism. What’s that all about?</td>
<td>Info-Request</td>
<td></td>
<td>Task</td>
</tr>
<tr>
<td>2</td>
<td>Student</td>
<td>It’s when you like stick together to things made of steel and iron and you and opposite sides repel and and what’s the other.</td>
<td>Answer</td>
<td>Task</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Tutor</td>
<td>What is up with the magnets sticking together at the ends or floating in the middle of this pencil?</td>
<td>Info-Request</td>
<td></td>
<td>Task</td>
</tr>
<tr>
<td>4</td>
<td>Student</td>
<td>They’re repelling.</td>
<td>Answer</td>
<td>Task</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.2: Example Tutorial Dialogue Annotated with DAMSL Acts

3.1.3.2 VERBMOBIL

VERBMOBIL (Bub and Schwinn, 1996) was an ambitious project to create a system for bidirectional, multilingual, simultaneous translation of spontaneous dialogues. Users of VERBMOBIL would engage in negotiation with another person using their native languages (German or Japanese) and VERBMOBIL would translate their speech into a language understood by both parties (English). The dialogue system as a whole assisted in the natural language understanding and translation. VERBMOBIL utilized dialogue acts for a variety of tasks including improving automatic speech
recognition, pruning dialogue decisions, and helping with plan recognition (Jekat et al., 1995).

At the highest level VERBMOBIL dialogue acts resemble the acts defined in DAMSL, however more descriptive acts were defined to better match the specific needs of VERBMOBIL’s tasks and application domains. In addition to defining acts to account for an utterance’s illocutionary force, Jekat et al. permitted finer-grained detail of DAMSL acts through the addition of subcategorization. In practice the subcategorization is a way to incorporate domain-specific semantic content into a subset of the acts. For example an accept act could be further refined into a more descriptive form from the appointment scheduling domain such as accept_date or accept_location.

Though subcategorization adds a level of information more aligned with the natural language interactions that arise in conversation, the domain-specific semantics of these labels restricts its utility for other tasks. Consequently, subcategorization was mostly dropped in the second edition of the VERBMOBIL dialogue act taxonomy (Alexandersson et al., 1998). Instead, Alexandersson et al. (1998) backed off to tags that capture only a small portion of the propositional content. Because VERBMOBIL-2 permitted multiple acts per utterance, the new taxonomy placed added emphasis on classifying acts by the kind of action they perform. In the new taxonomy, acts listed under the CONTROL_DIALOGUE category were concerned with managing the flow of communication. The function of acts in the MANAGE_TASK dealt with managing the opening and closing of tasks and topics. Lastly PROMOTE_TASK included everything concerned with solving or performing a task.

Although VERBMOBIL targets a set of concerns vastly different from those needed for modeling tutorial dialogue, it has implemented several attractive features that extend beyond high-level dialogue act descriptions. VERBMOBIL’s inclusion of mechanisms to account for task-relevant propositional content and its emphasis on separation of action and meaning have influenced the design of the DISCUSS taxonomy for tutorial dialogue.

3.1.3.3 Dynamic Interpretation Theory

Dynamic Interpretation Theory (DIT) (Bunt, 1997, 2000) was designed to model the flow and exchange of information between speakers during a task or goal oriented dialogue. Like earlier work
by Allen and Perrault (1980) and Cohen and Perrault (1979), DIT considers dialogue acts in light of their “context-changing effects and their corresponding dynamic meaning.” However, DIT largely regards planning and reasoning as separate from the dialogue model and instead focuses on creating a dialogue act taxonomy rich enough to represent subtleties in the dialogue state. Like DAMSL and other taxonomies, DIT and its successor DIT++ (Bunt, 2009; Bunt et al., 2010) endeavor to provide a multi-dimensional taxonomy suitable for many genres and domains. While creation of DAMSL is geared primarily toward use in dialogue analysis, DIT aspires to provide an annotation layer suitable for dialogue act generation. Consequently this much expanded taxonomy included features extending beyond the bounds of speech act theory. Of particular note were the addition of a subset of rhetorical structure theory (RST)-like roles (Mann and Matthiesson, 1989; Mann and Thompson, 1986, 1988), which were used as properties for DIT’s inform and as subclasses for the question act. Though they were not designed with tutorial dialogue in mind, they provide an additional richness unaccounted for in DAMSL and VERBMOBIL. Furthermore DIT’s division of dialogue control acts and task-oriented acts provides a straightforward framework for considering an utterance’s multiplicity of meaning. These additions of semantics and task-oriented details inspired much of the design of the DISCUSS taxonomy.

3.1.4 Tutorial Act and Question Type Taxonomies

In addition to the linguistically motivated work in speech and dialogue acts, there exists a large body of work directed at classifying and organizing behavior in an educational context. Unlike the more general dialogue acts, these works add much more detail to specific phenomena and behaviors that are crucial to learning.

3.1.4.1 Tutorial Act Annotation

Tsovaltzi and Karagjosova (2004) and Buckley and Wolska (2008) adapted DAMSL for tutoring by adding the more education-specific teaching task, solving task, and task progress dimensions. The creators of tutoring-DAMSL argued that these additional dimensions and the dialogue acts they
contained are useful for disambiguating moves for structured tasks including solving mathematics problems. However the omission of more specific question types does not address DAMSL’s shortcomings for more Socratic-style dialogues.

The DISCOUNT tutorial dialogue scheme (Pilkington, 1999) was designed to support two primary tasks: detailed analysis of educational discourse and natural language generation within automatic tutoring systems. DISCOUNT combined dialogue moves with aspects of Rhetorical Structure Theory (Mann and Thompson, 1988; Mann and Matthiesson, 1989; Mann and Thompson, 1986). Taken together, DISCOUNT dialogue moves, rhetorical function, and rhetorical predicates provide a picture of an utterance’s pragmatic interpretation and its surface form. However DISCOUNT by itself provides no account for the underlying semantics (events and entities) found within an utterance, and thus does not have a way to tie the utterance back to the goals of a lesson. The dialogue act, rhetorical form, and predicate type layers in DISCUSS are based in part on the divisions described by DISCOUNT.

3.1.4.2 Question Type Taxonomies

While it is important to have a breadth of move types in characterizing dialogue moves, it is often important to have depth to match the domain. In tutoring, the bulk of actions center on asking questions and providing answers. Detailed characterization of tutor and student behavior requires more than a shallow account of dialogue interactions. Assessing the educational merit of a tutorial dialogue needs a representation that can account for the variation in tutors’ questions and students’ answers.

Bloom’s taxonomy of education objectives (1956) is a foundational work in both the learning sciences and in intelligent tutoring systems research. Though it is not explicitly a taxonomy of question types, the hierarchy of skills in his cognitive domain provides a scale from which to gauge a question’s level of assessment, and many of the verbs associated with each level have direct correlates in many question type taxonomies found in the literature.
Graesser and Person (1994) defined a taxonomy to investigate the nature of question asking during tutoring. This taxonomy expanded both on the education-focused work of Bloom and the question classification schemes from research in artificial intelligence (Lehnert, 1978). At the highest level this scheme split questions into long and short answer questions. Question types within this taxonomy included categories such as Verification, Quantification, Definition, Comparison and Consequence.

More recent efforts have used bottom-up, data-driven approaches to create a taxonomy of questions appropriate for the task of automatic question generation (Nielsen et al., 2008; Boyer et al., 2009b). Nielsen et al.’s taxonomy synthesizes many of the features from past taxonomies (Bloom, 1956; Lehnert, 1978; Collins, 1985; Graesser and Person, 1994).

### 3.2 Applications and Techniques

#### 3.2.1 Dialogue Acts and Natural Language Processing

Despite the wide array of dialogue act taxonomies in the literature, dialogue acts themselves have seen limited use in actual natural language processing tasks. Much of the focus has been on the task of tagging utterances with dialogue acts (Levin et al., 1999; Webb et al., 2005; Louwerve and Crossley, 2006a; Surendran and Levow, 2006; Lan et al., 2008b) or in dialogue act prediction (Core, 1998; Geertzen, 2009). To date the most successful applications for dialogue act representations are in improving automatic speech recognition (Engel et al., 1995; Jurafsky et al., 1997; Stolcke et al., 2000) or in constraining options for dialogue planning (Engel et al., 1995; Reithinger and Maier, 1995).

#### 3.2.2 Tutorial Dialogue Analysis

There is a wide body of work that have analyzed corpora of both human-human and human-computer tutorial dialogue to identify features of the dialogue that correlate with learning. The majority of these analyses are based on metrics from coding schemes for dialogue moves and question
types along with shallower features of the dialogue.

Numerous studies have found strong correlation between the length of student turns and increased learning (Rosé et al., 2003; Core et al., 2003; Litman et al., 2004; Litman and Forbes-Riley, 2006; Litman et al., 2009), however the cause of this correlation is still up for debate. Many attribute this correlation to the common assumption that self-explanation promotes learning (Chi et al., 1989, 1994). Rosé et al. (2003) found that although closed-form questions could produce long answers, on average, open-ended questions produced longer responses. Core et al. (2003) hypothesized that student initiative could play a role, however they were unable to detect any correlation.

Using more sophisticated coding of dialogue acts on corpora of human-computer and human-human tutoring, Litman and Forbes-Riley (2006) found a positive correlation between questions requiring deeper answers and increased learning gains. However, they found little cross corpora overlap in the specific unigram and bigram patterns that led to learning.

Litman et al. (2009) applied a dialogue act coding scheme to corpora across a range of domains, modalities, and tutor types to see if prior findings generalized across environments. The investigation focused on the relationship between the “contentfulness” of student utterances and learning. They accomplished this by tagging words as “Domain Concepts” and then extracted the number of content words and number of content-rich turns as well as simple counts of linguistic units such as words and turns. To assist in this domain word matching, they employed common natural language techniques including stemming, stop-lists. Like with previous studies, they found that both longer turns and content-rich turns correlated with larger learning gains.

Moving towards more automatic techniques, Jeon and Azevedo (2008) utilized Coh-Metrix (Graesser et al., 2004) to compare the level of cohesion, coherence and readability between students that did and did not achieve large learning gains through tutoring. Coh-Metrix approximates cohesion between adjacent sentences using LSA (Landauer et al., 1998) similarity scores. Other metrics included simple measures such as number of turns, number of sentences, and number of words as well as more lexically motivated features like incidence of causal verbs and causal connectives.

Boyer et al. (2009a,c) trained hidden Markov models (HMM) on sequences of dialogue acts to
create a model of dialogue. Based on qualitative assessment Boyer et al. argued that the HMM hidden states correspond to the dialogue modes and strategies used by expert tutors described in Cade et al. (2008). She then included the presence of these modes as features in a classifier for predicting the potential learning gains from a dialogue session (Boyer, 2010).

These previous works in tutorial dialogue analysis have helped to both to confirm existing and discover new observations about the nature of tutoring. However, the use of dialogue acts for tutorial dialogue analysis has primarily made use of shallower, coarse-grained acts. There has been little work that ties the dialogue back to the semantics of the tutoring task. Consequently they are not detailed enough to give a true account of the student-tutor interactions. The DISCUSS dialogue move taxonomy detailed in the next chapter builds upon this existing body of research to drive towards providing a fuller account of the interactions found in tutorial dialogue and other task-oriented conversations.
DISCUSS: A multidimensional dialogue move taxonomy

The Dialogue Scheme for Unifying Speech and Semantics (DISCUSS) is a multifaceted dialogue move taxonomy intended to capture both the pragmatic and semantic interpretations of an utterance. A DISCUSS move is a tuple composed of values from four dimensions: Dialogue Act, Rhetorical Form, Predicate Type, and Semantic Roles. Together these dimensions convey the communicative action, surface form, and meaning of an utterance independent of the original utterance text.

DISCUSS is designed to serve as an intermediate representation for use in a variety of dialogue related tasks including dialogue move selection, automatic question generation, and analysis of dialogue sessions. In addition to these goals, DISCUSS desiderata include a taxonomic design that is descriptive, expandable, and domain (or curriculum) independent. The DISCUSS tags were developed from both bottom-up study of conversations in the MyST Wizard-of-Oz (WOZ) data and from synthesis of the dialogue act and tutorial move taxonomies detailed in Chapter 3. In particular, DISCUSS draws inspiration from the work of Pilkington (1999) and Nielsen et al. (2008). A complete listing of all the DISCUSS moves and dimensions can be found in the annotation guidelines detailed in Appendix A.

The rest of this chapter will motivate DISCUSS’ design through a dissection of its dimensions. It will start with overview of the broader move categories (Section 4.1), and then will progress

---

1 Parts of this chapter were adapted from DISCUSS: A dialogue move taxonomy layered over semantic representations, In Proceedings of the 9th International Conference of Computational Semantics (IWCS2011) (Becker et al., 2011).
to descriptions of the semantic role and predicate type in the section on semantic dimensions (Section 4.2) and the dialogue act and rhetorical form in the section on pragmatic dimensions (Section 4.3).

Notation: Throughout the remainder of this thesis DISCUSS moves are denoted using the following syntax:

Utterance text . . .

(Semantic Role).[domain/frame specific role].(role text)

Dialogue Act/Rhetorical Form/Predicate Type

Utterance text will be captured in an italic font; semantic roles will be denoted in violet text enclosed in angle brackets ⟨⟩; dialogue acts will be in blue text, rhetorical forms in green text and predicate types in red text. Together they will form a tuple delimited by a slash (/). Concrete examples of DISCUSS annotation can be seen in Figures 4.3 to 4.6.

4.1 Move Categories

DISCUSS moves are dictated by the dialogue act dimension and may belong to one of three broad categories: Dialogue Control, Information Exchange, and Attention Management. Dialogue Control moves are largely concerned with maintaining and enabling the flow of information. This includes dialogue acts such as Acknowledge, Open, Close, Repeat, and RequestRepeat. The Information Exchange moves relay content (often lesson-specific) between speakers using moves such as Assert, Ask, Answer, Mark, Revoice. For tutorial dialogue the bulk of student-tutor interactions reside in this category. Lastly, Attention Management moves indicate how a speaker exercises initiative over other speakers or topics. Dialogue acts found in the attention category are Focus, Defer, Elicit, and Direct.

Move categories are implicit to the dialogue act, consequently annotators tagging utterances with DISCUSS labels do not need to mark the move category. Instead, the move category should be used for considering the multiple actions an utterance may perform. Consider the example
in Figure 4.1 the tutor’s move covers all three categories by simultaneously providing feedback, directing a student to interact with a visual, and asking a follow-up question.

\[
\begin{align*}
T: \textit{Good Job! Try playing with the objects in this picture. What’s that all about?} \\
\text{Feedback/Positive/None (Dialogue Control)} \\
\text{Direct/Task/Visual (AttentionManagement)} \\
\text{Ask/Describe/Visual (Information Exchange)}
\end{align*}
\]

Figure 4.1: DISCUSS Move Categories Example: This illustrates how a single utterance can contain DISCUSS tuples from all three move categories.

4.2 Semantic Dimensions

The semantic dimensions define the objects, events, properties and relations contained within an utterance. The semantic roles at the lowest level of the DISCUSS hierarchy directly capture the propositional entities. Predicate types summarize the interactions between all of the semantic roles found within an utterance. Another way to think about the predicate type is as a more abstract label for frame-like constructions.

4.2.1 Semantic Roles

The MyST system models a lesson’s key concepts as propositions which are realized as semantic frames. These frames form the backbone of the MyST natural language understanding unit, and they are codified into the top-level nodes for semantic grammars used by the Phoenix parser (Ward, 1994). An example concept/frame and Phoenix parse is shown in Figure 4.2. When using MyST, student responses are parsed with a manually written semantic grammar allowing paraphrases and partial-paraphrases of the frame description to fill the frame elements. While the goal of tutoring is to elicit a response that entails and fills all facets in a frame, the frame-element level semantic decomposition provides tutors with follow-up points when student answers are incomplete or incorrect.

Although these semantic frames form the basis of MyST dialogue, to allow for domain-
Figure 4.2: Semantic Role Example: Example MyST semantic frame, student utterance and corresponding Phoenix parser output

independent generalization of dialogue behavior across a wide range of subjects requires a higher-level semantic representation. VerbNet (Schuler, 2005) was chosen as a starting point for defining a set of semantic roles because of its intuitive balance between descriptiveness and portability. While a majority of the labels were used as is, some roles needed to be modified and others needed to be added to properly cover the set of concepts used by MyST. For example, many key concepts that express proportionality relationships can not be easily represented using predicate argument structure. Consequently, the decomposition for the concept “The strength of the force decreases as distance increases.” is a cause (the distance relationship) and an effect (the force relationship). An analysis of the frames and concepts resulted in the addition of the part, whole, configuration, term, and definition roles. Ambiguity regarding attribute led us to split it into two new roles: predicate-attribute and entity-attribute. Lastly, the catch-all keyword label was added to reflect terms that may relate to the proposition, but are not part of the core representation.

4.2.2 Predicate Type

Simply knowing an utterance’s propositional content is insufficient for inferring what was stated. Consider the two exchanges in Figure 4.3. The mixture of semantic roles in both students’ responses are identical. Additionally, one can not differentiate between the exchanges based solely on dialogue act or rhetorical form. We need additional information to know that the tutor in the first scenario seeks to elicit discussion about observations while the conversation in the second
scenario focuses on procedures. Beyond simply describing the process, one can also imagine such information would be useful for identifying communication breakdowns. For example, responding with a description of a procedure to a request about a process may indicate that the student did not understand the question or that the student is unwilling or unable to address the question.

To address this need, the **Predicate Type** dimension was created based partly on the tutoring question types described by Graesser and Person (1994) and Nielsen et al. (2008) and partly on the rhetorical predicates used in the DISCOUNT (Pilkington, 1999) scheme. While DISCOUNT included discourse relations in the set of predicate types, predicate types are restricted to those that encapsulate or summarize the collection of semantic roles in an utterance. Example predicate types include *Function*, *Procedure*, *Route*, *Observation* and *Purpose*.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Utterance</th>
<th>DISCUSS tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tutor:</td>
<td>Tell me about what's going on here in this picture.</td>
<td>Ask/Describe/Observation</td>
</tr>
<tr>
<td>Student:</td>
<td>The wires connect the battery and the light bulb and then light bulb lights up. (Instrument).wires (Predicate).connect (Theme1).battery (Theme2).light bulb (Effect).bulb lights up</td>
<td>Answer/Describe/Observation</td>
</tr>
<tr>
<td>Tutor:</td>
<td>Tell me about how you got the bulb to light up.</td>
<td>Ask/Describe/Procedure</td>
</tr>
<tr>
<td>Student:</td>
<td>To make the light go we connected the wires to the battery and the bulb. (Effect).light go (Predicate).connected (Instrument).wires (Theme1).battery (Theme2).bulb</td>
<td>Answer/Describe/Procedure</td>
</tr>
</tbody>
</table>

Figure 4.3: Predicate Type Example: Two scenarios with identical dialogue acts, rhetorical forms, and semantic role labels. The predicate type (shown in boldfaced, underlined red text) serves as a tool for disambiguation.

### 4.3 Pragmatic Dimensions

The pragmatic dimensions are composed of the dialogue act dimension and the rhetorical form dimension. The dialogue act expresses the communicative action of a move and is the most general dimension in DISCUSS. The rhetorical form expresses the function and attributes of the utterance’s surface realization and can be thought of as refining the intent of the coarser dialogue act.
4.3.1 Dialogue Act Dimension

The dialogue act dimension is the top-level dimension in DISCUSS with the values of all other dimensions depending on the value of this dimension. Like with the majority of dialogue act taxonomies, DISCUSS dialogue acts have a grounding in speech act theory with a focus on what action the utterance performs. While most of the dialogue acts in the communicative and informational move categories have direct corollaries to those found in other taxonomies like DIT++ or DAMSL, they were further supplemented with two grounding acts commonly used in Questioning the Author instruction: mark and revoice. In marking, the tutor highlights parts of the student’s language to emphasize important points and to steer the conversation towards key concepts. Revoicing serves a similar purpose, but instead of highlighting, the tutor rephrases student speech to clarify ideas they may have been struggling with. Examples of each of these acts are shown in Figures 4.4 and 4.5.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Utterance</th>
<th>DISCUSS tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student:</td>
<td>that when you stick a magnet to a rusty nail and then you stick it</td>
<td>Answer/Describe/Process</td>
</tr>
<tr>
<td></td>
<td>to a paper clip it sticks</td>
<td></td>
</tr>
<tr>
<td>Tutor:</td>
<td>I think I heard you say something about magnets sticking or attracting.</td>
<td>Mark/None/None</td>
</tr>
<tr>
<td></td>
<td>Tell me more about that.</td>
<td>Ask/Elaborate/Process</td>
</tr>
</tbody>
</table>

Figure 4.4: Marking Example

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Utterance</th>
<th>DISCUSS tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student:</td>
<td>well when you scrub the the paperclip to the magnet the paperclip</td>
<td>Answer/Describe/Process</td>
</tr>
<tr>
<td></td>
<td>is starting to begin to be a magnet</td>
<td></td>
</tr>
<tr>
<td>Tutor:</td>
<td>Very good, so if the magnet gets close to the paperclip it picks it up.</td>
<td>Feedback/Positive/None</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Revoice/None/None</td>
</tr>
</tbody>
</table>

Figure 4.5: Revoice Example

Dialogue acts in the dialogue control category also reflect many of the actions regularly seen in tutorial dialogue. Focus and Defer acts are often used to move to or away from lesson-specific topics. With MyST dialogues Direct is typically used with utterances that give instructions related
to the multimedia (e.g. “Click on the box” or “Look at this animation.”).

### 4.3.2 Rhetorical Form

The DISCUSS **rhetorical form** dimension provides another mechanism for differentiating between utterances with identical semantic content. While the dialogue act dimension is useful for providing an utterance’s pragmatic interpretation and for determining what sequences are licensed, by itself it provides no indication of how a speaker is advancing the topic under discussion. From a language generation perspective, additional information is needed to create an utterance’s surface form. Dialogue acts provide a coarse approximation, but a rhetorical relation is needed to tie the high level pragmatics to the underlying semantics. Consider the three transactions in Figure 4.6. The semantic parses in all three scenarios would be identical, however the tutor’s questions and the resulting student response all serve very different functions. In the first the tutor is asking for a description, in the second identification, and in the third justification. Selection of the DISCUSS rhetorical forms found in the Informational move category were inspired by the sixteen top-level tags used in Rhetorical Structure Theory (RST) (Mann and Thompson, 1986, 1988; Mann and Matthiesson, 1989). Similar to how RST uses a rhetorical relation to link clauses and to show the development of an argument, DISCUSS uses the rhetorical form to relate turns to other turns and to show development of the dialogue and tutoring strategy.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Utterance</th>
<th>DISCUSS tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Tutor:</td>
<td><em>Can you describe what is going on with the battery?</em></td>
<td>Ask/Describe/Visual</td>
</tr>
<tr>
<td>Student:</td>
<td><em>The battery is putting out electricity.</em></td>
<td>Answer/Describe/Process</td>
</tr>
<tr>
<td>2 Tutor:</td>
<td><em>Can you tell which one is the battery?</em></td>
<td>Ask/Identify/Entity</td>
</tr>
<tr>
<td>Student:</td>
<td><em>The battery is the one putting out the electricity.</em></td>
<td>Answer/Identify/Entity</td>
</tr>
<tr>
<td>3 Tutor:</td>
<td><em>Ok, why do you think that one is the battery?</em></td>
<td>Ask/Justify/Entity</td>
</tr>
<tr>
<td>Student:</td>
<td><em>That one’s the battery cause it’s putting out electricity.</em></td>
<td>Answer/Justify/Entity</td>
</tr>
</tbody>
</table>

Figure 4.6: Rhetorical Form Example: These three question/answer pairs highlight the relationship between an utterance’s surface form and its rhetorical form.
4.4 Discussion

This chapter introduced the DISCUSS dialogue move taxonomy, and annotation scheme that overlays dialogue act and rhetorical annotation over semantic representations. The design of this system draws from past work in dialogue and discourse annotation, semantic representations, and tutorial move taxonomies. Together this combination of pragmatic and semantic layers provides an intermediate representation suitable for analyzing and automating interactions in a complex task-oriented domain like tutorial dialogue. Subsequent chapters will detail experiments that demonstrate the utility of DISCUSS for various dialogue-related tasks.
Chapter 5

Annotating Tutorial Dialogues with DISCUSS

While a dialogue representation can stand on its own, the true test of its merit comes in applying its principles to real data. An annotation study using the MyST Wizard-of-Oz (WOZ) dialogues was carried out to create a corpus of DISCUSS tagged dialogues. This study served to investigate, test, and refine the DISCUSS taxonomy, and the completed corpus is used as a resource for building the dialogue models described in subsequent chapters.

5.1 Annotators

Two undergraduate-trained linguists served as the project annotators for this corpus annotation study. Training in the DISCUSS taxonomy was an iterative process with incremental refinements to the definitions and labels. The final guidelines are detailed in Appendix A.

This project initially started with one annotator and a second was hired to allow for assessment of inter-annotator reliability and to help complete annotation.

5.2 Annotation Tool

To facilitate annotation of MyST Wizard-of-Oz dialogues with DISCUSS, the author of this thesis wrote a full-featured web application called DISCUSSed (DISCUSS editor). Annotators using this tool can login and load any of their assigned dialogues at any time. This enables them to not only work incrementally on a given dialogue, it provides a resource for them to cross-reference and/or revisit past judgments.
Upon opening a dialogue session in DISCUSSed, the annotator is shown a screen with two tabs labeled “Propositions” and “Turns”. The “Propositions” tab (Figure 5.1) lists the propositions or learning goals for the particular FOSS unit targeted by this tutoring session. Below each proposition are the semantic frame elements used by the Phoenix dialogue manager in parsing student utterances. This tab is provided as a resource to aid the annotator in understanding the underlying goals and motivations driving the tutors’ decisions.

The bulk of the DISCUSS annotation work occurs in the “Turns” tab (Figure 5.2). Within this screen, annotators are presented with the full dialogue history. Next to each turn is a timestamp, speaker ID, as well as the proposition the tutor is currently trying to elicit from the student. Tutor turns can be composed of three types of commands synth, flash and clear_screen. A synth command shows the words output by the text-to-speech (TTS) system, flash commands list the
Figure 5.2: Screenshot of DISCUSSed Annotation Tool Turns Tab.

visual presented to the student during the turn, and clear_screen simply means any visual on display was removed. Student turns show the transcribed text and a Phoenix-produced semantic parse of the utterance.

DISCUSS annotation actually occurred at the segment level. For tutor turns, segmentation was performed automatically with each flash and synth command forming a distinct segment. For example, a tutor’s turn with text “synth(Let’s look at something); flash(parallel_circuit); synth(Tell me what’s going on here.)” would have three distinct segments. It is the annotator’s job to label
the synth segments with DISCUSS tuples. With student turns, there is no punctuation or easy way to segment turns, so tutors were asked to use DISCUSSed to segment long terms into segments more conducive to labeling.

To label moves with DISCUSS dialogue acts, rhetorical forms, and predicate types, annotators selected items from drop-down boxes. The DISCUSSed tool was populated with knowledge of the taxonomy, which allowed the contents of the rhetorical form and predicate type boxes to change dynamically depending on the annotations selected in other layers. This functionality helped to prevent annotators from labeling turns with nonsensical tuple such as Feedback/Define/Entity. Annotators were free to add as many DISCUSS tuples to a segment as they thought necessary.

5.3 Semantic Roles Annotation

Rather than manually tagging each and every utterance in the corpus with VerbNet labels, a mapping layer was hand-crafted to append semantic role labels to output from the Phoenix parser output. Figure 5.3 shows the role mappings for the frame in Figure 4.2 along with a post-mapping parse. In the future, with enough training data it could be possible to replace the hand-tuned semantic grammars with a statistically trained semantic role labeler.

<table>
<thead>
<tr>
<th>Frame:</th>
<th>DirectionOfFlow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description:</td>
<td>Electricity flows from the negative side of the battery to the positive side</td>
</tr>
<tr>
<td>(Theme):</td>
<td>[Electricity]</td>
</tr>
<tr>
<td>(Predicate):</td>
<td>[Flows]</td>
</tr>
<tr>
<td>(Source):</td>
<td>[FromTerminal]</td>
</tr>
<tr>
<td>(Destination):</td>
<td>[ToTerminal]</td>
</tr>
<tr>
<td>Utterance:</td>
<td>The energy it it goes from the minus side to the plus</td>
</tr>
</tbody>
</table>
| Parse:          | DirectionOfFlow:(Theme):[Electricity].energy (Predicate):[Flow].goes  
|                 | (Source):[FromTerminal],[Negative].minus 
|                 | (Destination):[ToTerminal],[Positive].plus |

Figure 5.3: Example mapping between MyST Semantic Frame Elements and VerbNet-style Semantic Role Labels along with a post-mapping Phoenix parse
5.4 Corpus Statistics

In total 119 transcripts covering ten investigations from Magnetism and Electricity and two from Measurement were used in this corpus annotation study. Altogether 5786 turns were tagged with DISCUSS labels. Of these, 2588 were student turns and 3198 were tutor turns. These transcripts represent dialogues from 14 different tutors and 52 different students.

<table>
<thead>
<tr>
<th>Dialogue Act</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tutor</td>
<td>Student</td>
</tr>
<tr>
<td>Acknowledge</td>
<td>70</td>
<td>44</td>
</tr>
<tr>
<td>Answer</td>
<td>26</td>
<td>2281</td>
</tr>
<tr>
<td>Apologize</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Ask</td>
<td>2049</td>
<td>2</td>
</tr>
<tr>
<td>Assert</td>
<td>457</td>
<td>8</td>
</tr>
<tr>
<td>Close</td>
<td>118</td>
<td>57</td>
</tr>
<tr>
<td>Defer</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Direct</td>
<td>550</td>
<td>0</td>
</tr>
<tr>
<td>Feedback</td>
<td>697</td>
<td>3</td>
</tr>
<tr>
<td>Focus</td>
<td>129</td>
<td>0</td>
</tr>
<tr>
<td>Hint</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>Mark</td>
<td>118</td>
<td>0</td>
</tr>
<tr>
<td>Metastatement</td>
<td>53</td>
<td>45</td>
</tr>
<tr>
<td>Open</td>
<td>103</td>
<td>117</td>
</tr>
<tr>
<td>Recall</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>Repeat</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>RequestRepeat</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Revoice</td>
<td>107</td>
<td>0</td>
</tr>
<tr>
<td>SignalNoUnderstanding</td>
<td>3</td>
<td>80</td>
</tr>
<tr>
<td>Thank</td>
<td>26</td>
<td>13</td>
</tr>
<tr>
<td>Uninterpretable</td>
<td>1</td>
<td>38</td>
</tr>
</tbody>
</table>

Table 5.1: DISCUSS Dialogue Act Frequencies
### Table 5.2: DISCUSS Rhetorical Form Frequencies

<table>
<thead>
<tr>
<th>Rhetorical Form</th>
<th>Frequency Tutor</th>
<th>Frequency Student</th>
<th>Relative Frequency Tutor</th>
<th>Relative Frequency Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attend</td>
<td>200</td>
<td>0</td>
<td>6.05%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Bye</td>
<td>113</td>
<td>55</td>
<td>3.42%</td>
<td>2.04%</td>
</tr>
<tr>
<td>Clarify</td>
<td>9</td>
<td>5</td>
<td>0.27%</td>
<td>0.19%</td>
</tr>
<tr>
<td>Compare</td>
<td>90</td>
<td>103</td>
<td>2.72%</td>
<td>3.81%</td>
</tr>
<tr>
<td>Confirm</td>
<td>29</td>
<td>23</td>
<td>0.88%</td>
<td>0.85%</td>
</tr>
<tr>
<td>Define</td>
<td>33</td>
<td>20</td>
<td>1.00%</td>
<td>0.74%</td>
</tr>
<tr>
<td>Describe</td>
<td>1289</td>
<td>1247</td>
<td>38.98%</td>
<td>46.19%</td>
</tr>
<tr>
<td>Elaborate</td>
<td>523</td>
<td>228</td>
<td>15.81%</td>
<td>8.44%</td>
</tr>
<tr>
<td>Greet</td>
<td>102</td>
<td>114</td>
<td>3.08%</td>
<td>4.22%</td>
</tr>
<tr>
<td>Identify</td>
<td>151</td>
<td>151</td>
<td>4.57%</td>
<td>5.59%</td>
</tr>
<tr>
<td>Justify</td>
<td>26</td>
<td>22</td>
<td>0.79%</td>
<td>0.81%</td>
</tr>
<tr>
<td>List</td>
<td>74</td>
<td>153</td>
<td>2.24%</td>
<td>5.67%</td>
</tr>
<tr>
<td>Meta</td>
<td>87</td>
<td>0</td>
<td>2.63%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Negative</td>
<td>1</td>
<td>1</td>
<td>0.03%</td>
<td>0.04%</td>
</tr>
<tr>
<td>Positive</td>
<td>696</td>
<td>2</td>
<td>21.05%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Predict</td>
<td>5</td>
<td>5</td>
<td>0.15%</td>
<td>0.19%</td>
</tr>
<tr>
<td>Quantify</td>
<td>28</td>
<td>35</td>
<td>0.85%</td>
<td>1.30%</td>
</tr>
<tr>
<td>Recap</td>
<td>279</td>
<td>0</td>
<td>8.44%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Select</td>
<td>6</td>
<td>6</td>
<td>0.18%</td>
<td>0.22%</td>
</tr>
<tr>
<td>Task</td>
<td>313</td>
<td>0</td>
<td>9.46%</td>
<td>0.00%</td>
</tr>
<tr>
<td>YesNo</td>
<td>37</td>
<td>36</td>
<td>1.12%</td>
<td>1.33%</td>
</tr>
</tbody>
</table>

### Table 5.3: DISCUSS Predicate Type Frequencies

<table>
<thead>
<tr>
<th>Predicate Type</th>
<th>Frequency Tutor</th>
<th>Frequency Student</th>
<th>Relative Frequency Tutor</th>
<th>Relative Frequency Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>AcceptRejectMaybe</td>
<td>7</td>
<td>22</td>
<td>0.21%</td>
<td>0.90%</td>
</tr>
<tr>
<td>Activity</td>
<td>117</td>
<td>65</td>
<td>3.54%</td>
<td>2.65%</td>
</tr>
<tr>
<td>Attribute</td>
<td>134</td>
<td>216</td>
<td>4.05%</td>
<td>8.82%</td>
</tr>
<tr>
<td>CausalRelation</td>
<td>419</td>
<td>562</td>
<td>12.67%</td>
<td>22.95%</td>
</tr>
<tr>
<td>Configuration</td>
<td>119</td>
<td>198</td>
<td>3.60%</td>
<td>8.08%</td>
</tr>
<tr>
<td>Duration</td>
<td>2</td>
<td>1</td>
<td>0.06%</td>
<td>0.04%</td>
</tr>
<tr>
<td>Entity</td>
<td>358</td>
<td>225</td>
<td>10.83%</td>
<td>9.19%</td>
</tr>
<tr>
<td>Experience</td>
<td>33</td>
<td>15</td>
<td>1.00%</td>
<td>0.61%</td>
</tr>
<tr>
<td>Function</td>
<td>148</td>
<td>164</td>
<td>4.48%</td>
<td>6.70%</td>
</tr>
<tr>
<td>Location</td>
<td>19</td>
<td>21</td>
<td>0.57%</td>
<td>0.86%</td>
</tr>
<tr>
<td>Observation</td>
<td>365</td>
<td>222</td>
<td>11.04%</td>
<td>9.06%</td>
</tr>
<tr>
<td>Procedure</td>
<td>59</td>
<td>64</td>
<td>1.78%</td>
<td>2.61%</td>
</tr>
<tr>
<td>Process</td>
<td>82</td>
<td>49</td>
<td>2.48%</td>
<td>2.00%</td>
</tr>
<tr>
<td>Proposition</td>
<td>478</td>
<td>5</td>
<td>14.45%</td>
<td>0.20%</td>
</tr>
<tr>
<td>Route</td>
<td>65</td>
<td>164</td>
<td>1.97%</td>
<td>6.70%</td>
</tr>
<tr>
<td>Topic</td>
<td>150</td>
<td>83</td>
<td>4.54%</td>
<td>3.39%</td>
</tr>
<tr>
<td>Visual</td>
<td>733</td>
<td>136</td>
<td>22.17%</td>
<td>5.55%</td>
</tr>
<tr>
<td>YesNoMaybe</td>
<td>16</td>
<td>35</td>
<td>0.48%</td>
<td>1.43%</td>
</tr>
</tbody>
</table>
5.5 Inter-Annotator Agreement

A reliability study using 15% of the 119 transcripts was conducted to assess inter-rater agreement of DISCUSS tagging. This consisted of 18 doubly-annotated transcripts comprised of 828 dialogue utterances.

Cohen’s Kappa (\(\kappa\)) (Carletta, 1996), a statistic for measuring reliability between raters on categorical items is used as the primary measure of inter-annotator agreement. Because DISCUSS permits multiple labels per instance, a \(\kappa\) value is computed for each label, which then contributes to the mean agreement for each DISCUSS dimension. Specifically the \(\kappa\) value is measuring how well the annotators agreed on the presence or absence of a specific label for a given utterance. The detailed breakdown of \(\kappa\) by label and DISCUSS dimension is shown in Tables 5.5 to 5.7. For most DISCUSS classes the majority classification is not present, consequently the raw percent agreement scores are much higher than that chance adjusted \(\kappa\) values.

To get an additional sense of agreement, two other metrics are used: exact agreement and partial agreement. For each of these metrics, each annotator’s annotations are treated as a per class bag-of-labels. For exact agreement, each annotator’s set of labels must match exactly to receive credit. Partial agreement is defined as the number of intersecting labels divided by the total number of unique labels. Together these statistics help to bound the reliability of the DISCUSS annotation. Table 5.4 lists all three metrics broken down by DISCUSS dimension. The \(\kappa\) values show fair agreement for the dialogue act and rhetorical form dimensions, whereas the predicate type shows more moderate agreement. This difference reflects the relative difficulty in labeling each dimension, and the agreement as a whole illustrates the open-endedness of the task. More importantly it sets an upper bound on how well a computer can do on this task, and gives confidence that automatic DISCUSS labeling is possible.
<table>
<thead>
<tr>
<th>Reliability Metric</th>
<th>DA</th>
<th>RF</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen’s Kappa</td>
<td>0.76</td>
<td>0.72</td>
<td>0.63</td>
</tr>
<tr>
<td>Exact Agreement</td>
<td>0.80</td>
<td>0.66</td>
<td>0.56</td>
</tr>
<tr>
<td>Partial Agreement</td>
<td>0.89</td>
<td>0.77</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Table 5.4: Summary of inter-annotator agreement by DISCUSS dimensions (DA=Dialogue Act, RF=Rhetorical Form, PT=Predicate Type)

<table>
<thead>
<tr>
<th>Dialogue Act</th>
<th>Kappa ($\kappa$)</th>
<th>% Agree</th>
<th>Num. Tagged Annotator 1</th>
<th>Num. Tagged Annotator 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledge</td>
<td>0.363</td>
<td>0.980</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>Answer</td>
<td>0.967</td>
<td>0.985</td>
<td>311</td>
<td>314</td>
</tr>
<tr>
<td>Apologize</td>
<td>0.799</td>
<td>0.998</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Ask</td>
<td>0.958</td>
<td>0.980</td>
<td>337</td>
<td>336</td>
</tr>
<tr>
<td>Assert</td>
<td>0.882</td>
<td>0.980</td>
<td>83</td>
<td>76</td>
</tr>
<tr>
<td>Close</td>
<td>0.914</td>
<td>0.995</td>
<td>26</td>
<td>22</td>
</tr>
<tr>
<td>Defer*</td>
<td>1.000</td>
<td>1.000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Direct</td>
<td>0.910</td>
<td>0.984</td>
<td>88</td>
<td>86</td>
</tr>
<tr>
<td>Elicit*</td>
<td>0.415</td>
<td>0.972</td>
<td>27</td>
<td>15</td>
</tr>
<tr>
<td>Feedback</td>
<td>0.673</td>
<td>0.933</td>
<td>126</td>
<td>69</td>
</tr>
<tr>
<td>Focus</td>
<td>0.745</td>
<td>0.982</td>
<td>31</td>
<td>30</td>
</tr>
<tr>
<td>Hint</td>
<td>0.666</td>
<td>0.998</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Mark</td>
<td>0.679</td>
<td>0.987</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>Metastatement</td>
<td>0.580</td>
<td>0.984</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Open</td>
<td>0.956</td>
<td>0.996</td>
<td>36</td>
<td>35</td>
</tr>
<tr>
<td>Recall</td>
<td>0.399</td>
<td>0.996</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Repeat</td>
<td>0.632</td>
<td>0.991</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>RequestRepeat*</td>
<td>0.000</td>
<td>0.996</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Revoice</td>
<td>0.767</td>
<td>0.988</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>SignalNoUnderstanding</td>
<td>0.908</td>
<td>0.998</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Thank</td>
<td>1.000</td>
<td>1.000</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Uninterpretable</td>
<td>0.780</td>
<td>0.994</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>All</td>
<td><strong>0.767</strong></td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 5.5: DISCUSS Dialogue Act Inter-Annotator Agreement. Items marked with an asterisk (*) are not included in the mean kappa value as they did not occur frequently enough in the doubly-adjudicated sample.
<table>
<thead>
<tr>
<th>Rhetorical Form</th>
<th>Kappa (κ)</th>
<th>% Agree</th>
<th>Num. Tagged Annotator 1</th>
<th>Num. Tagged Annotator 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attend</td>
<td>0.827</td>
<td>0.985</td>
<td>38</td>
<td>41</td>
</tr>
<tr>
<td>Bye</td>
<td>0.934</td>
<td>0.996</td>
<td>25</td>
<td>22</td>
</tr>
<tr>
<td>Clarify</td>
<td>1.000</td>
<td>1.000</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Compare</td>
<td>0.642</td>
<td>0.969</td>
<td>31</td>
<td>45</td>
</tr>
<tr>
<td>Confirm</td>
<td>0.724</td>
<td>0.993</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>Define</td>
<td>0.278</td>
<td>0.982</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>Describe</td>
<td>0.762</td>
<td>0.881</td>
<td>428</td>
<td>469</td>
</tr>
<tr>
<td>Elaborate</td>
<td>0.652</td>
<td>0.933</td>
<td>108</td>
<td>75</td>
</tr>
<tr>
<td>Greet</td>
<td>0.956</td>
<td>0.996</td>
<td>36</td>
<td>35</td>
</tr>
<tr>
<td>Identify</td>
<td>0.589</td>
<td>0.944</td>
<td>60</td>
<td>66</td>
</tr>
<tr>
<td>Justify</td>
<td>0.749</td>
<td>0.998</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>List</td>
<td>0.310</td>
<td>0.957</td>
<td>40</td>
<td>15</td>
</tr>
<tr>
<td>Meta</td>
<td>0.703</td>
<td>0.994</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Negative*</td>
<td>1.000</td>
<td>1.000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Positive</td>
<td>0.673</td>
<td>0.933</td>
<td>126</td>
<td>69</td>
</tr>
<tr>
<td>Predict</td>
<td>0.666</td>
<td>0.999</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Quantify</td>
<td>0.836</td>
<td>0.994</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>Recap</td>
<td>0.826</td>
<td>0.982</td>
<td>46</td>
<td>45</td>
</tr>
<tr>
<td>Select*</td>
<td>1.000</td>
<td>1.000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Task</td>
<td>0.890</td>
<td>0.988</td>
<td>50</td>
<td>46</td>
</tr>
<tr>
<td>YesNo</td>
<td>0.746</td>
<td>0.991</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td><strong>0.724</strong></td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 5.6: DISCUSS Rhetorical Form Inter-Annotator Agreement. Items marked with an asterisk (*) are not included in the mean kappa value as they did not occur frequently enough in the doubly-adjudicated sample.
<table>
<thead>
<tr>
<th>Predicate Type</th>
<th>Kappa (κ)</th>
<th>% Agree</th>
<th>Num. Tagged Annotator 1</th>
<th>Num. Tagged Annotator 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AcceptRejectMaybe</td>
<td>0.543</td>
<td>0.994</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Activity</td>
<td>0.838</td>
<td>0.987</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>Attribute</td>
<td>0.706</td>
<td>0.966</td>
<td>53</td>
<td>52</td>
</tr>
<tr>
<td>CausalRelation</td>
<td>0.641</td>
<td>0.898</td>
<td>175</td>
<td>114</td>
</tr>
<tr>
<td>Configuration</td>
<td>0.531</td>
<td>0.941</td>
<td>45</td>
<td>69</td>
</tr>
<tr>
<td>Duration</td>
<td>1.000</td>
<td>1.000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Entity</td>
<td>0.544</td>
<td>0.897</td>
<td>105</td>
<td>117</td>
</tr>
<tr>
<td>Experience</td>
<td>0.908</td>
<td>0.998</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Function</td>
<td>0.666</td>
<td>0.948</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>Location*</td>
<td>0.000</td>
<td>0.991</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Observation</td>
<td>0.544</td>
<td>0.931</td>
<td>70</td>
<td>71</td>
</tr>
<tr>
<td>Procedure</td>
<td>0.687</td>
<td>0.984</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td>Process</td>
<td>0.279</td>
<td>0.927</td>
<td>27</td>
<td>63</td>
</tr>
<tr>
<td>Proposition</td>
<td>0.581</td>
<td>0.937</td>
<td>85</td>
<td>55</td>
</tr>
<tr>
<td>Route</td>
<td>0.796</td>
<td>0.978</td>
<td>52</td>
<td>47</td>
</tr>
<tr>
<td>Topic</td>
<td>0.509</td>
<td>0.974</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td>Visual</td>
<td>0.844</td>
<td>0.958</td>
<td>146</td>
<td>130</td>
</tr>
<tr>
<td>YesNoMaybe</td>
<td>0.495</td>
<td>0.986</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td><strong>0.632</strong></td>
<td></td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 5.7: DISCUSS Predicate Type Inter-Annnotator Agreement. Items marked with an asterisk (*) are not included in the mean kappa value as they did not occur frequently enough in the doubly-adjudicated sample.
Analysis of unseen tutorial dialogues requires the ability to automatically label utterances with the DISCUSS dialogue acts, rhetorical forms, and predicate types. For an intelligent tutoring system and many WOZ-style interactions the tutor’s moves come from a preexisting pool of prompts. Because the tutoring moves are a fixed part of the system, the relative cost of manual DISCUSS annotation is small as it can be amortized across all future sessions in the form of a table lookup.

Conversely, student responses are the most dynamic aspect of open-ended tutorial dialogue. This variability eliminates the possibility of a simple lookup, and dictates the need for more intelligent approaches to labeling student moves. The existing corpus of DISCUSS annotated WOZ dialogues provides the necessary data to train DISCUSS utterance taggers. Specifically, the pairing of utterance segments with DISCUSS labels provides the examples necessary for machine learning to induce a function that maps features of the utterance to labels. The remainder of this chapter details how to approach DISCUSS tagging as a supervised learning task, and it reports on the performance of systems trained on annotation from the DISCUSS WOZ corpus. The results demonstrate the feasibility of this task while also giving insight into the relative difficulty of classifying different DISCUSS labels and dimensions. Detailed analysis of the results are also used to refine the DISCUSS taxonomy to further improve the automatic utterance classifiers.
6.1 Related Work

This work draws inspiration from a large body of existing work in dialogue act recognition and classification. Many of the techniques described later in this chapter have been successfully employed in these other systems. Furthermore, past research has helped to frame the experimental design and evaluation framework used for automatic DISCUSS labeling.

In recent years, much of the effort in dialogue act recognition has largely focused on using supervised machine learning to identify patterns in the utterances that correlate with different dialogue act categories. Stolcke, et al. Stolcke et al. (2000) utilized a combination of lexical, prosodic and syntactic cues for identifying dialogue acts. Louwerse and Crossley (2006b) make use of a number of n-gram features for this task. Klüwer et al. (2010) employed dependency grammars and VerbNet (Schuler, 2005) frames to identify specific syntactic and semantic structures for classification. Verbree et al. (2006) showed that dialogue classification can perform as high as 80% on corpora with as many as 80,000 utterances.

In the intelligent tutoring systems literature and tutorial dialogue literature, there are two notable examples of dialogue act classification for student utterances. Olney et al. (2003) used hand-crafted finite state grammars and other shallow NLP techniques to obtain a 0.98 weighted F-measure. Boyer et al. (2010) obtained 68.2% labeling accuracy using a combination of supervised classification and Hidden Markov Modeling.

There has been less emphasis on classifying dialogue acts in multiple dimensions, especially for tutorial dialogues. Though modeling multiple dimensions introduces additional complexity, it offers the potential for a more descriptive, meaningful account of the dialogue. DAMSL (Core and Allen, 1997) was among the first taxonomies designed with multiple levels; however most work in DAMSL act classification flattens the taxonomy into a set of mutually exclusive tags. Clark and Popescu-Belis (2004) achieved 77.9% accuracy in initial experiments with a corpus tagged with the MALTUS dialogue act taxonomy.
6.2 DISCUSS Classification

Unlike most other dialogue move taxonomies, DISCUSS utilizes multiple dimensions and permits multiple tags per utterance. Consequently, the task of labeling dialogues with DISCUSS is not simply a matter of choosing a single category. Much of the existing work assumes a single label per utterance. Moreover most taxonomies limit the number of possible dialogue acts to 10 different categories.

Enumeration of all combinations of DISCUSS dialogue acts, rhetorical forms, and predicate types would yield hundreds of different categories. This creates a sparsity of positive examples and poses a challenge for training a labeling system. To make the problem more tractable the strict constraint of treating the dialogue act, rhetorical form, and predicate type as a single tuple is relaxed. Instead the utterance’s labels can be considered a bag of DISCUSS labels. This parallels the bag approach used to compute inter-annotator agreement. More importantly, recasting of the representation provides a more straightforward way to learn utterance classification models. Instead of learning to select a single DISCUSS tuple, the task shifts outwards with a separate binary classifier for each DISCUSS label.

6.3 Selected DISCUSS Labels

In a MyST tutoring session, no restrictions are placed on the students’ speech or language. From a dialogue-representation perspective, students are capable of taking action in line with any of the moves from the DISCUSS taxonomy. However the tag frequencies (Tables 5.1 to 5.3) observed in the Wizard-of-Oz corpus show student utterances cover only a subset of the available DISCUSS labels. To ensure there are sufficient data for the learning algorithm, labels with a relative frequency less than 0.5% within the set of student turns are not included. This threshold eliminated 12 dialogue act labels, 9 rhetorical form labels, and 2 predicate type labels. The instances tagged with remaining labels (listed below) are used to train the DISCUSS utterance classifiers, which are later employed as part of the dialogue data mining study described in Chapter 8. The set of DISCUSS
tags used for classifying student utterances are as follows:

**Dialogue Acts** Acknowledge, Answer, Close, Metastatement, SignalNoUnderstanding, Thanks, and Uninterpretable

**Rhetorical Form** Bye, Compare, Confirm, Define, Describe, Elaborate, Greet, Identify, Justify, List, Quantify, YesNo

**Predicate Type** AcceptRejectMaybe, Activity, Attribute, CausalRelation, Configuration, Entity, Experience, Function, Location, Observation, Procedure, Process, Route, Topic, Visual, YesNoMaybe

6.3.1 **Merging DISCUSS Labels**

While the DISCUSS labels selected in the previous section meet the minimum occurrence threshold, sparsity is still a barrier to system performance (see Section 6.7). This is further exaggerated when split across multiple folds for cross-validation. As observed in both the results and the corpus statistics, there is at least an order of magnitude difference between the number of positive (tag present) and negative (tag not present) instances for all labels. In practice this compounds the already difficult challenge of managing the trade-off between precision and recall.

Closer inspection of classifier errors (outlined below in Section 6.8) revealed that in many cases pairs of labels were indistinguishable from an annotation perspective. This ambiguity leads directly to classifier confusion. For example a *Describe* rhetorical form classifier may incorrectly identify an utterance with the *Elaborate* tag as being a *Describe*. Similarly, there is large overlap in the lexicon, construction, and context for utterances with the *SignalNoUnderstanding* and *Uninterpretable* dialogue acts.

Merging labels is one way to reduce the effects of sparsity and ambiguity, while still retaining the original annotation. This merging makes for more clearly distinguishable categories and reduces noise along label boundaries. Palmer et al. (2006) used analyses of manual tagging for word sense
disambiguation (WSD) to identify groups of fine-grained senses that could be collapsed into coarse-grained senses. They found this merging of classes provided a more meaningful unit of analysis and allowed systems to achieve state-of-the-art results in the WSD task (Dligach and Palmer, 2008).

To obtain more reliable dialogue move classifiers for later use in dialogue analysis, commonly confused DISCUSS tags were merged into new tags, which were subsequently used to train and evaluate a new set of classifiers. The merged classes and their constituent labels are listed in Table 6.1. More justification for the merged categories can be found in the detailed classifier error analysis provided in Appendix B.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Merged Label</th>
<th>Original Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue Acts</td>
<td>MergedNoUnderstanding</td>
<td>SignalNoUnderstanding</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Uninterpretable</td>
</tr>
<tr>
<td>Rhetorical Forms</td>
<td>MergedDescribe</td>
<td>Define</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Describe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elaborate</td>
</tr>
<tr>
<td></td>
<td>MergedYesNo</td>
<td>Confirm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YesNo</td>
</tr>
<tr>
<td>Predicate Types</td>
<td>MergedTopic</td>
<td>Activity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experience</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic</td>
</tr>
<tr>
<td></td>
<td>MergedCausalRelation</td>
<td>CausalRelation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Process</td>
</tr>
<tr>
<td></td>
<td>MergedRoute</td>
<td>Location</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Route</td>
</tr>
<tr>
<td></td>
<td>MergedVisual</td>
<td>Observation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Visual</td>
</tr>
<tr>
<td></td>
<td>MergedYesNoMaybe</td>
<td>AcceptRejectMaybe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YesNoMaybe</td>
</tr>
</tbody>
</table>

Table 6.1: Merged DISCUSS Labels
6.4 Preprocessing

In addition to the raw text and DISCUSS labels, standard NLP preprocessing is needed to enable extraction of features. Both tutor and student utterance segments were processed using readily available components wrapped in the ClearTK (Ogren et al., 2008) statistical natural language processing framework. The preprocessing pipeline consisted of the ClearTK tokenizer, the OpenNLP (“Apache OpenNLP Development Community”, 2010) sentence segmenter and part-of-speech (POS) tagger and the Clear Morphological Analyzer (Choi and Nicolov, 2009).

6.5 Features

Classification of DISCUSS tags utilizes features extracted from the student’s utterance and the preceding dialogue context. These features are motivated by three observations:

(1) Cue phrases are explicit indicators of the structure of discourse (Hirschberg and Litman, 1993).

(2) Entrainment, the process of automatic alignment between dialogue partners is a key factor in facilitating successful tutorial dialogue (Ward et al., 2011a).

(3) Dialogue context is predictive of subsequent dialogue moves (Stolcke et al., 2000).

The feature categories listed in Table 6.2 correspond to each of these intuitions, and the feature descriptions detail the specific value extracted into the feature vector.
<table>
<thead>
<tr>
<th>Category</th>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterance</td>
<td>BoW1</td>
<td>utterance bag-of-words (unigram)</td>
</tr>
<tr>
<td></td>
<td>BoW2</td>
<td>utterance bag-of-words (bigram)</td>
</tr>
<tr>
<td></td>
<td>BoPOS1</td>
<td>utterance bag-of-POS tags (unigram)</td>
</tr>
<tr>
<td></td>
<td>BoPOS2</td>
<td>utterance bag-of-POS tags (bigram)</td>
</tr>
<tr>
<td>Number words</td>
<td></td>
<td>utterance contains a number word</td>
</tr>
<tr>
<td>Entrainment</td>
<td>OverlapWord1</td>
<td>word unigram overlap (stud. utt., prev. tutor utt.)</td>
</tr>
<tr>
<td></td>
<td>OverlapWord2</td>
<td>word bigram overlap (stud. utt., prev. tutor utt.)</td>
</tr>
<tr>
<td></td>
<td>OverlapLem1</td>
<td>lemma unigram overlap (stud. utt., prev. tutor utt.)</td>
</tr>
<tr>
<td></td>
<td>OverlapLem2</td>
<td>lemma bigram overlap (stud. utt., prev. tutor utt.)</td>
</tr>
<tr>
<td></td>
<td>OverlapPOS1</td>
<td>POS tag unigram overlap (stud. utt., prev. tutor utt.)</td>
</tr>
<tr>
<td></td>
<td>OverlapPOS2</td>
<td>POS tag bigram overlap (stud. utt., prev. tutor utt.)</td>
</tr>
<tr>
<td>Dialogue</td>
<td>prevBoDA</td>
<td>previous utt. bag-of-DISCUSS Dialogue Acts</td>
</tr>
<tr>
<td>Context</td>
<td>prevBoRF</td>
<td>previous utt. bag-of-DISCUSS Rhetorical Forms</td>
</tr>
<tr>
<td></td>
<td>prevBoPT</td>
<td>previous utt. bag-of-DISCUSS Predicate Types</td>
</tr>
</tbody>
</table>

Table 6.2: DISCUSS Classifier Features

6.6 Model Training, Parameter Selection, and Evaluation

With features defined and extracted, the next step in building a DISCUSS classifier is learning a model. All DISCUSS tagging models were trained with the ClearTK (Ogren et al., 2008) wrapper for OpenNLP’s Maximum Entropy classifier\(^1\) using the default model parameters. Maximum Entropy (aka MaxEnt or regularized-Logistic Regression) is a statistical modeling framework that integrates information from multiple information sources for classification (Berger et al., 1996). This ability to combine evidence from multiple, non-independent sources is well suited to natural language processing tasks (Ratnaparkhi, 1998), and its minimal number of parameters simplifies training.

Along with the default MaxEnt parameters, two other training parameters were added to deal with the uneven distribution of positive and negative instances caused by label sparsity:

\[ p_{\text{neg}} \] - Specifies the probability of keeping a negative example. During instance creation time a random number is selected and if it is less than or equal to \( p_{\text{neg}} \) the instance is included in the set of training examples.

\(^1\) http://opennlp.apache.org/
freqpos - Indicates the number of times to duplicate a positive example within the set of training examples.

Without the $p_{neg}$ and $freq_{pos}$ parameters to govern the inclusion of an instance in the training set, machine learning algorithms including MaxEnt are prone to overfitting with a heavy bias towards the majority class. For many DISCUSS labels, this would result in a classifier that only predicts **false** or **not present**. Model parameters were selected via grid search over 10-fold cross validation, with highest $F_1$-score serving as the selection criterion. The grid search ranges were $p_{neg} = \{0.1, 0.25, 0.5, 1.0\}$ and $freq_{pos} = \{1, 2, 5, 10\}$, and the parameters for each DISCUSS label’s top-performing model are shown in Table 6.3.

### 6.6.1 Evaluation

Because DISCUSS labeling is performed with multiple binary classifiers, evaluation is carried out on a per-label basis. Though accuracy is a common metric for evaluating binary classifiers, the skewed ratio between positive and negative examples reduces the utility and informativeness of this measure. Specifically, this imbalance means that near perfect accuracies can be achieved by always guessing the majority label (**not present** for most DISCUSS tags). More meaningful measures reflect the classifiers’ sensitivity and specificity in different operating conditions.

In particular, evaluation uses three primary measures **Precision**, **Recall**, and $F_1$-score. Precision (Equation (6.1)) is the proportion of utterances labeled by the system as positive (having a Dialogue Act, Rhetorical Form, or Predicate Type), that are true positives (correct). Recall (Equation (6.2)) is the proportion of actual positives which are correctly identified as such. $F_1$-score (Equation (6.3)) is the harmonic mean of precision and recall and can be interpreted as a weighted average of the precision and recall. All three scores are calculated using counts of **True Positives** (TP), **True Negatives** (TN), **False Positives** (FP), and **False Negatives** (FN).
<table>
<thead>
<tr>
<th>Dialogue Acts</th>
<th>Label</th>
<th>$p_{neg}$</th>
<th>$freq_{pos}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledge</td>
<td>0.1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Answer</td>
<td>1.0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Close</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Metastatement</td>
<td>0.5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Open</td>
<td>0.25</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>SignalNoUnderstanding</td>
<td>1.0</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Thank</td>
<td>0.25</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Uninterpretable</td>
<td>0.1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>MergedNoUnderstanding</td>
<td>0.5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhetorical Form</td>
<td>Bye</td>
<td>1.0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Compare</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Confirm</td>
<td>1.0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Define</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Describe</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Elaborate</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Greet</td>
<td>1.0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Identify</td>
<td>1.0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Justify</td>
<td>0.25</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>List</td>
<td>0.25</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Quantify</td>
<td>0.5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>YesNo</td>
<td>1.0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>MergedDescribe</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>MergedYesNo</td>
<td>1.0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicate Type</td>
<td>AcceptRejectMaybe</td>
<td>0.5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Activity</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Attribute</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>CausalRelation</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Configuration</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Entity</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Experience</td>
<td>0.1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Function</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Observation</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Procedure</td>
<td>0.25</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Process</td>
<td>0.1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Route</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Topic</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Visual</td>
<td>0.25</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>YesNoMaybe</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>MergedTopic</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>MergedCausalRelation</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>MergedRoute</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>MergedVisual</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>MergedYesNoMaybe</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 6.3: DISCUSS Model Parameters
\[
Precision = \frac{TP}{TP + FP} \tag{6.1}
\]
\[
Recall = \frac{TP}{TP + FN} \tag{6.2}
\]
\[
F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.3}
\]

### 6.7 Results

The average $F_1$-score for all non-merged dialogue act classifiers was 0.710, rhetorical form classifiers 0.585, and predicate type classifiers 0.484. Adjusting these scores based on relative frequencies yields weighted average $F_1$-scores of 0.935 for dialogue acts, 0.704 for rhetorical forms, and 0.530 for predicate types. Tables 6.4 to 6.6 further breakdown these scores for each DISCUSS tag as well as for the merged tags.

<table>
<thead>
<tr>
<th>Dialogue Act</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Acc.</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledge</td>
<td>44</td>
<td>2617</td>
<td>34</td>
<td>9</td>
<td>0.984</td>
<td>0.564</td>
<td>0.830</td>
<td>0.672</td>
</tr>
<tr>
<td>Answer</td>
<td>2226</td>
<td>363</td>
<td>101</td>
<td>14</td>
<td>0.957</td>
<td>0.957</td>
<td>0.994</td>
<td>0.975</td>
</tr>
<tr>
<td>Close</td>
<td>57</td>
<td>2632</td>
<td>4</td>
<td>11</td>
<td>0.994</td>
<td>0.934</td>
<td>0.838</td>
<td>0.884</td>
</tr>
<tr>
<td>Metastatement</td>
<td>25</td>
<td>2573</td>
<td>77</td>
<td>29</td>
<td>0.961</td>
<td>0.245</td>
<td>0.463</td>
<td>0.321</td>
</tr>
<tr>
<td>Open</td>
<td>120</td>
<td>2572</td>
<td>4</td>
<td>8</td>
<td>0.996</td>
<td>0.968</td>
<td>0.938</td>
<td>0.952</td>
</tr>
<tr>
<td>SignalNoUnderstanding</td>
<td>85</td>
<td>2593</td>
<td>13</td>
<td>13</td>
<td>0.990</td>
<td>0.867</td>
<td>0.867</td>
<td>0.867</td>
</tr>
<tr>
<td>Thank</td>
<td>11</td>
<td>2688</td>
<td>0</td>
<td>5</td>
<td>0.998</td>
<td>1.000</td>
<td>0.688</td>
<td>0.815</td>
</tr>
<tr>
<td>Uninterpretable</td>
<td>11</td>
<td>2606</td>
<td>57</td>
<td>30</td>
<td>0.968</td>
<td>0.162</td>
<td>0.268</td>
<td>0.202</td>
</tr>
<tr>
<td>MergedNoUnderstanding</td>
<td>87</td>
<td>2559</td>
<td>6</td>
<td>52</td>
<td>0.979</td>
<td>0.935</td>
<td>0.626</td>
<td>0.750</td>
</tr>
</tbody>
</table>

Table 6.4: DISCUSS Dialogue Act Classifier Performance

TP=number of True Positives, TN=number of True Negatives, FP=number of False Positives, FN=number of False Negatives, Acc=Accuracy, P=Precision, R=Recall, F=$F_1$-score. Precision is the proportion of utterances labeled by the system as positive (having a Dialogue Act), that are true positives (correct). Recall is the proportion of actual positives instances which are correctly identified as positive.

Dialogue Acts starting with “Merged” are derived by combining multiple dialogue acts into one class.
<table>
<thead>
<tr>
<th>Rhetorical Form</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Acc</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bye</td>
<td>59</td>
<td>2627</td>
<td>11</td>
<td>7</td>
<td>0.993</td>
<td>0.843</td>
<td>0.894</td>
<td>0.868</td>
</tr>
<tr>
<td>Compare</td>
<td>64</td>
<td>2516</td>
<td>63</td>
<td>61</td>
<td>0.954</td>
<td>0.504</td>
<td>0.512</td>
<td>0.508</td>
</tr>
<tr>
<td>Confirm</td>
<td>10</td>
<td>2679</td>
<td>2</td>
<td>13</td>
<td>0.994</td>
<td>0.833</td>
<td>0.435</td>
<td>0.571</td>
</tr>
<tr>
<td>Define</td>
<td>7</td>
<td>2659</td>
<td>10</td>
<td>28</td>
<td>0.986</td>
<td>0.412</td>
<td>0.200</td>
<td>0.269</td>
</tr>
<tr>
<td>Describe</td>
<td>1452</td>
<td>580</td>
<td>621</td>
<td>51</td>
<td>0.751</td>
<td>0.700</td>
<td>0.966</td>
<td>0.812</td>
</tr>
<tr>
<td>Elaborate</td>
<td>121</td>
<td>2203</td>
<td>258</td>
<td>122</td>
<td>0.859</td>
<td>0.319</td>
<td>0.498</td>
<td>0.389</td>
</tr>
<tr>
<td>Greet</td>
<td>119</td>
<td>2572</td>
<td>7</td>
<td>6</td>
<td>0.995</td>
<td>0.944</td>
<td>0.952</td>
<td>0.948</td>
</tr>
<tr>
<td>Identify</td>
<td>82</td>
<td>2421</td>
<td>115</td>
<td>86</td>
<td>0.926</td>
<td>0.416</td>
<td>0.488</td>
<td>0.449</td>
</tr>
<tr>
<td>Justify</td>
<td>8</td>
<td>2664</td>
<td>8</td>
<td>24</td>
<td>0.988</td>
<td>0.500</td>
<td>0.250</td>
<td>0.333</td>
</tr>
<tr>
<td>List</td>
<td>106</td>
<td>2426</td>
<td>108</td>
<td>64</td>
<td>0.936</td>
<td>0.495</td>
<td>0.624</td>
<td>0.552</td>
</tr>
<tr>
<td>Quantify</td>
<td>20</td>
<td>2661</td>
<td>3</td>
<td>20</td>
<td>0.991</td>
<td>0.870</td>
<td>0.500</td>
<td>0.635</td>
</tr>
<tr>
<td>YesNo</td>
<td>23</td>
<td>2660</td>
<td>3</td>
<td>18</td>
<td>0.992</td>
<td>0.885</td>
<td>0.561</td>
<td>0.687</td>
</tr>
<tr>
<td>MergedDescribe</td>
<td>1640</td>
<td>648</td>
<td>305</td>
<td>111</td>
<td>0.846</td>
<td>0.843</td>
<td>0.937</td>
<td>0.887</td>
</tr>
<tr>
<td>MergedYesNo</td>
<td>38</td>
<td>2632</td>
<td>8</td>
<td>26</td>
<td>0.987</td>
<td>0.826</td>
<td>0.594</td>
<td>0.691</td>
</tr>
</tbody>
</table>

Table 6.5: DISCUSS Rhetorical Form Classifier Performance
TP=number of True Positives, TN=number of True Negatives, FP=number of False Positives, FN=number of False Negatives, Acc=Accuracy, P=Precision, R=Recall, F=$F_1$-score. Precision is the proportion of utterances labeled by the system as positive (having a Rhetorical Form), that are true positives (correct). Recall is the proportion of actual positive instances which are correctly identified as positive.

Rhetorical Form starting with “Merged” are derived by combining multiple rhetorical forms into one class.
<table>
<thead>
<tr>
<th>Predicate Type</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Acc.</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>AcceptRejectMaybe</td>
<td>10</td>
<td>2676</td>
<td>6</td>
<td>12</td>
<td>0.993</td>
<td>0.625</td>
<td>0.455</td>
<td>0.526</td>
</tr>
<tr>
<td>Activity</td>
<td>51</td>
<td>2568</td>
<td>64</td>
<td>21</td>
<td>0.969</td>
<td>0.443</td>
<td>0.708</td>
<td>0.545</td>
</tr>
<tr>
<td>Attribute</td>
<td>183</td>
<td>2278</td>
<td>155</td>
<td>88</td>
<td>0.910</td>
<td>0.541</td>
<td>0.675</td>
<td>0.601</td>
</tr>
<tr>
<td>CausalRelation</td>
<td>505</td>
<td>1636</td>
<td>462</td>
<td>101</td>
<td>0.792</td>
<td>0.522</td>
<td>0.833</td>
<td>0.642</td>
</tr>
<tr>
<td>Configuration</td>
<td>91</td>
<td>2410</td>
<td>89</td>
<td>114</td>
<td>0.925</td>
<td>0.506</td>
<td>0.444</td>
<td>0.473</td>
</tr>
<tr>
<td>Entity</td>
<td>169</td>
<td>2199</td>
<td>253</td>
<td>83</td>
<td>0.876</td>
<td>0.400</td>
<td>0.671</td>
<td>0.501</td>
</tr>
<tr>
<td>Experience</td>
<td>13</td>
<td>2562</td>
<td>109</td>
<td>20</td>
<td>0.952</td>
<td>0.107</td>
<td>0.394</td>
<td>0.168</td>
</tr>
<tr>
<td>Function</td>
<td>160</td>
<td>2265</td>
<td>213</td>
<td>66</td>
<td>0.897</td>
<td>0.429</td>
<td>0.708</td>
<td>0.534</td>
</tr>
<tr>
<td>Location</td>
<td>8</td>
<td>2676</td>
<td>7</td>
<td>13</td>
<td>0.993</td>
<td>0.533</td>
<td>0.381</td>
<td>0.444</td>
</tr>
<tr>
<td>Observation</td>
<td>103</td>
<td>2294</td>
<td>176</td>
<td>131</td>
<td>0.886</td>
<td>0.369</td>
<td>0.440</td>
<td>0.402</td>
</tr>
<tr>
<td>Procedure</td>
<td>44</td>
<td>2532</td>
<td>60</td>
<td>68</td>
<td>0.953</td>
<td>0.423</td>
<td>0.393</td>
<td>0.407</td>
</tr>
<tr>
<td>Process</td>
<td>8</td>
<td>2583</td>
<td>64</td>
<td>49</td>
<td>0.958</td>
<td>0.111</td>
<td>0.140</td>
<td>0.124</td>
</tr>
<tr>
<td>Route</td>
<td>116</td>
<td>2453</td>
<td>87</td>
<td>48</td>
<td>0.950</td>
<td>0.571</td>
<td>0.707</td>
<td>0.632</td>
</tr>
<tr>
<td>Topic</td>
<td>75</td>
<td>2560</td>
<td>48</td>
<td>21</td>
<td>0.974</td>
<td>0.610</td>
<td>0.781</td>
<td>0.685</td>
</tr>
<tr>
<td>Visual</td>
<td>60</td>
<td>2419</td>
<td>133</td>
<td>92</td>
<td>0.917</td>
<td>0.311</td>
<td>0.395</td>
<td>0.348</td>
</tr>
<tr>
<td>YesNoMaybe</td>
<td>23</td>
<td>2663</td>
<td>1</td>
<td>17</td>
<td>0.993</td>
<td>0.958</td>
<td>0.575</td>
<td>0.719</td>
</tr>
<tr>
<td>MergedTopic</td>
<td>138</td>
<td>2467</td>
<td>38</td>
<td>61</td>
<td>0.963</td>
<td>0.784</td>
<td>0.693</td>
<td>0.736</td>
</tr>
<tr>
<td>MergedCausalRelation</td>
<td>558</td>
<td>1551</td>
<td>499</td>
<td>96</td>
<td>0.780</td>
<td>0.528</td>
<td>0.853</td>
<td>0.652</td>
</tr>
<tr>
<td>MergedRoute</td>
<td>135</td>
<td>2434</td>
<td>85</td>
<td>50</td>
<td>0.950</td>
<td>0.614</td>
<td>0.730</td>
<td>0.667</td>
</tr>
<tr>
<td>MergedVisual</td>
<td>283</td>
<td>1865</td>
<td>455</td>
<td>101</td>
<td>0.794</td>
<td>0.383</td>
<td>0.737</td>
<td>0.504</td>
</tr>
<tr>
<td>MergedYesNoMaybe</td>
<td>37</td>
<td>2632</td>
<td>10</td>
<td>25</td>
<td>0.987</td>
<td>0.787</td>
<td>0.597</td>
<td>0.679</td>
</tr>
</tbody>
</table>

Table 6.6: DISCUSS Predicate Type Classifier Performance
TP=number of True Positives, TN=number of True Negatives, FP=number of False Positives, FN=number of False Negatives, Acc=Accuracy, P=Precision, R=Recall, F=F-score. Precision is the proportion of utterances labeled by the system as positive (having a Predicate Type), that are true positives (correct). Recall is the proportion of actual positive instances which are correctly identified as positive.

Predicate types starting with “Merged” are derived by aggregating multiple predicate types into one class.
6.7.1 Merged Labels Results

Precision-Recall curves were plotted for each of the “Merged” labels along with their constituent tags (Figures 6.1 to 6.8). Each point represents the cross-validation recall and precision achieved with a given set of model training parameters. These curves help to compare the relative improvement achieved by merging labels. Ideal performance is in the upper, right-hand corner of the graph at precision=1.0 and recall=1.0.

![Figure 6.1: MergedNoUnderstanding Dialogue Act Precision-Recall Curve](image)
Figure 6.2: MergedDescribe Rhetorical Form Precision-Recall Curve

Figure 6.3: MergedYesNo Rhetorical Form Precision-Recall Curve
Figure 6.4: *MergedTopic* Predicate Type Precision-Recall Curve

Figure 6.5: *MergedCausalRelation* Predicate Type Precision-Recall Curve
Figure 6.6: MergedRoute Predicate Type Precision-Recall Curve

Figure 6.7: MergedVisual Predicate Type Precision-Recall Curve
6.7.2 Discussion

The relative difficulty of classifying dialogue acts, rhetorical forms, and predicate types directly follows from the inter-annotator agreement statistics. Average system $F_1$-scores approach the Cohen’s Kappa values. These results are encouraging, and they demonstrate the feasibility of automatically tagging student speech with labels from the DISCUSS taxonomy. Pairing the results with a detailed error analysis in Section 6.8 reveals that when classifiers do not agree with the gold-standard annotation they are not necessarily incorrect, and may in fact be behaving correctly. Error analysis also helped to identify which labels led to classifier confusion and provided insight into which DISCUSS labels should be merged.

Merging labels greatly improves tagging robustness, with an average gain in $F_1$-score of 0.19 points. Closer inspection of the precision-recall curves for the merged labels gives a clearer understanding of the effects of the merging. Figure 6.1 shows that the benefits of merging SignalNoUnderstanding with Uninterpretable are not entirely clear. On the one hand it allows the data tagged Uninterpretable to be used, on the other hand it degrades performance on a reli-
able *SignalNoUnderstanding* classifier. The curves for *MergedDescribe* (Figure 6.2), *MergedTopic* (Figure 6.4), *MergedCausalRelation* (Figure 6.5) and *MergedVisual* (Figure 6.7) give clear visual confirmation that folding the tags together produces a more useful utterance classifier. Conversely, the precision-recall plots for *MergedYesNo* (Figure 6.3) and *MergedYesNoMaybe* (Figure 6.8) and *MergedRoute* (Figure 6.6) tell a less decisive story. Though the merged label classifiers exhibit better performance their curves do not clearly dominate the individual classifiers.

The plots for the *MergedYesNo* rhetorical form classifier and the *MergedYesNoMaybe* predicate type classifier exhibit nearly identical morphology and overlap. This suggests that there is little information to be gained by having both a rhetorical form and predicate type for yes-no utterances. Differentiation would require partitioning the predicate type into distinct *Yes*, *No* and *Maybe* tags.

### 6.8 Error Analysis

In the interest of space and continuity, this section will highlight the findings uncovered by the error analysis. For a more thorough analysis of each of the individual DISCUSS classifiers, please refer to Appendix B.

Analysis of the classifier errors helped to identify four common sources of error in classifying student utterances with DISCUSS labels:

1. Lexical overfitting
2. Context sensitive errors
3. Annotator error
4. Evaluation limitations

#### 6.8.1 Lexical overfitting

In most natural language processing tasks, the lexical features are the most predictive, and dialogue move labeling is no exception. While the words of an utterance are the most discriminative feature, they can also bias the machine learning algorithm to overfitting – especially when the corpus
is small and splintered across multiple domains like it is with the WOZ-DISCUSS corpus. Lexical overfitting was the most common and pervasive cause of classifier error.

For example the Attribute predicate type classifier was prone to committing false positive errors when the utterance contained lesson-specific words such as “aluminum”, “iron” or “metal”. Utterances (a) and (b) were incorrectly identified as having predicate type Attribute when they were actually tagged with the CausalRelation and Observation tags.

a) **Student:** and it connects to anything that has iron in it
   
   Answer/Describe/CausalRelation

b) **Student:** it didn’t pick the alumninum nail

   Answer/Describe/Observation

Lexical overfitting was not only a source of false positive errors, but it could also cause errors of omission (false negative errors). The name “Marnie” (the name of the tutor within MyST) was predominantly spoken at the tail end of the session when the student was saying goodbye. Even though the student’s speech in utterance (c) contained the very indicative “hello”, the presence of “Marnie” was enough to confuse the Open dialogue act classifier.

c) **Student:** hello marnie

Syntactic features such as dependency relations could help to avoid these over-generalizations, but at a cost of much more training data. Similarly features based on semantic parses could help to recognize different usage patterns or constructions that are indicative of dialogue act, rhetorical form or predicate type.

### 6.8.2 Context sensitive errors

Lexically identical utterances carry different actions and different semantics depending on context. This shifting meaning caused difficulties for several classifiers, especially when the utterances were short. The student responses listed in examples (a)-(d) can all be interpreted as acknowledgment,
agreement, confirmation, or a yes-no response. This confusion led to false positive errors for any confusable classifier.

a) Student: okay
b) Student: well
c) Student: yeah
d) Student: mm hmm yes

In other cases, the usual lexical cues were not present in the utterance, and the DISCUSS label could only be inferred from context. In example (e) the rhetorical form for both the tutor and student utterances should be Compare, however the absence of prototypical Compare cues like comparative and quantifiers resulted in a false negative.

e) Tutor: This is a parallel circuit. How does this compare to other circuits you have made?

Student: the wires don’t touch each other and they and there’s two wires connected to the positive and negative side unlike some other times there’s only one connected to the positive and one connected to the negative

Though the DISCUSS classifiers capture context with bag-of-DISCUSS tags from the previous utterance, these errors highlight the need for additional dialogue context features. The most likely next step would be the inclusion of features based on the transition probabilities between DISCUSS tags.

6.8.3 Annotator error

As the inter-annotator agreement statistics would indicate, the scope and size of the DISCUSS taxonomy cause inherently more noise than typically found in other more well defined NLP tasks. Many of the errors came about because the distinction between tags is more of a continuum than a hard line. In example (a) the DISCUSS classifier considered the student response to have predicate
type Function; however the annotator interpreted this statement as a CausalRelation. This kind of annotation error can be mitigated with further refinement of the annotation guidelines.

a) Tutor: Tell me more about what the wires do in a circuit

(Ask/Describe/Function)

Student: they bring the energy from the battery to the motor which powers the motor

(Answer/Describe/CausalRelation)

6.8.4 Evaluation limitations

Though the DISCUSS classifiers were evaluated using widely accepted approaches, these measures may be more rigid and more pessimistic than is warranted by the annotation task. For many utterances there are several equally valid interpretations, which can lead to very different annotations. Furthermore, human annotators typically assigned only a single tuple to an utterance, when there may have been others that applied. Examples (a)-(c) show three utterances and their gold standard annotations. These were all scored as false positives for the Compare rhetorical form classifier. However, one could argue that both the decisions made by the annotators and the decision made by the Compare classifier are correct as there are both contextual and lexical features that support the decision. Closer examination shows that the Quantify and MergedDescribe are making the correct true positive decision suggesting that a simultaneous evaluation that assigns partial credit may give a more well-rounded view of classifier errors.
a) Student:  *there are two circuits one on the right and one on the left*  
   (Answer/Quantify/Entity)
b) Student:  *and the circuit on the on the left side is the light bulb is brighter because there’s only one light bulb and on the on the right side there’s two light bulbs so it’s*  
   (Answer/MergedDescribe/MergedVisual)
c) Student:  *some of the straws were longer than the other ones and some of them were shorter than the other*  
   (Answer/MergedDescribe/Attribute)

6.9 Conclusions

This chapter introduced the task of automatic labeling of student utterances in tutorial dialogues with moves from the DISCUSS taxonomy. More importantly this work provides a starting point for exploring the practical and technical limitations associated with recognizing dialogue acts, rhetorical forms and predicate types in unseen speech. The methods detailed above provide a straightforward, tractable mechanism for dealing with complex, multi-dimensional, multi-label dialogue move taxonomies. Recasting what is traditionally a strict, multi-class classification task as a series of binary decisions circumvents the hard decision of choosing a single best label, and allows for finer tuning of tagging behavior.

Using a set of standard features for dialogue move classification in conjunction with DISCUSS-specific features yielded a set of promising results. The frequency-adjusted $F_1$-scores of 0.935 for dialogue acts, 0.704 for rhetorical forms, and 0.530 for predicate types are on par with corpus inter-annotator agreement statistics. Training and evaluation of these classifiers yielded a detailed error analysis, which shed light on the common problems and pitfalls associated with this task. These analyses also provide guidance to further refine the DISCUSS taxonomy and to improve the annotation process. Merging similar or ambiguous DISCUSS labels helped to reduce noise and sparsity in the training data, and ultimately gave a boost in classification performance. The refined
taxonomy and collection of DISCUSS classifiers are used in Chapter 8 to investigate the role of DISCUSS in characterizing the potential for learning in tutorial dialogues.
Chapter 7

Question Ranking and Selection in Context

An overarching goal of this thesis is to improve the dialogue capabilities in intelligent tutoring systems. This chapter focuses on the crucial subtask of selecting follow-up questions within the context of a tutorial dialogue. Although asking questions is only a subset of the overall tutoring process, it is still a complex process that requires understanding of the dialogue state, the student’s ability, and the learning goals. The challenge in this task is not simply to pick a context-relevant question, but to prioritize those that also encourage self-expression and stimulate learning and learner interest.

This work frames question selection as a task of scoring and ranking candidate questions for a specific point in the tutorial dialogue. Since dialogue is a dynamic process with multiple correct possibilities, the potential moves and questions used in this study are not restricted only to those found in the MyST-WOZ corpus described in Chapter 5. Instead this work explores the possibilities that stem from the question “What if a fully automatic question generation system existed?”. This is accomplished through the use of candidate questions hand-authored for each dialogue context.

To investigate the mechanisms involved in ranking follow-up questions, these questions are paired with judgments of quality from experienced human tutors. Features extracted from the questions’ surface form, and underlying DISCUSS dialogue representation are embedded in machine learning classification algorithms to ultimately learn a function for ranking the appropriateness of questions.

1 Parts of this chapter were adapted from Learning to Tutor Like a Tutor: Question Ranking in Context, In proceedings of the 11th International Conference on Intelligent Tutoring Systems (Becker et al., 2012a) and Question Ranking and Selection in Tutorial Dialogues, In proceedings of the 7th workshop on Building Educational Applications using NLP (Becker et al., 2012b).
for specific points in a dialogue.

These results show promise with the best question ranking models exhibiting performance on par with experienced human tutors. Furthermore, training models on judgments collected from individual judges (tutors) helps to shed light on the differences in pedagogical style and questioning tactics – even among tutors with similar training and backgrounds. The experiments and results detailed below provide three major contributions toward the larger goal of enabling computers to learn tutorial dialogue policies directly from human examples. First they demonstrate the utility and importance of rich dialogue representations, such as DISCUSS, for modeling decision making in task-oriented dialogues. Second, they provide a framework for learning behavior from Wizard-of-Oz data. Lastly, the question ranking task gives a scaffold for evaluating and learning behaviors using fully-automatic question generation.

7.1 Connections to Prior Work

Learning tutorial dialogue policies from corpora is a growing area of research in NLP and ITS. Existing work has made use of hidden Markov models (Boyer et al., 2009a) and reinforcement learning (Chi et al., 2010, 2008) to discover tutoring strategies optimized to maximize learning gains; however, much of this work assumes there is only one correct behavior, and the additional complexity required to model individual tutoring styles would require much more data. This work adopts an approach similar to Ai and Litman (2008) who utilize ranking to predict human judgments of simulated dialogue quality.

There is also an expanding body of work that applies ranking algorithms toward the task of question generation (QG) using approaches such as over-generation-and-ranking (Heilman and Smith, 2010a), language model ranking (Yao, 2010), and heuristics-based ranking (Agarwal and Mannem, 2011). While the focus of these efforts centers on issues of grammaticality, fluency, and content selection for automatic creation of standalone questions, the experiments described in this chapter shift focus towards the higher level task of choosing context appropriate questions. The work presented in this chapter merges aspects of these QG approaches with the sentence planning
tradition from natural language generation (Walker et al., 2001; Rambow et al., 2001). In sentence planning the goal is to select lexico-structural resources that encode communicative action. Rather than selecting representations, this system uses them directly as part of the feature space for learning functions to rank the questions’ actual surface form realization.

Previous work in categorizing dialogue acts and questions for tutoring (Graesser and Person, 1994; Core and Allen, 1997; Pilkington, 1999) has helped to shed light on the nature of interactions between tutors and students. Corpora tagged with dialogue and tutoring acts have been used to explore the correlation between tutoring moves and learning (Jackson et al., 2004; Litman and Forbes-Riley, 2006) as well as specific behaviors such as when to ask “why” questions (Rosé et al., 2003), provide hints (Tsoveltzi and Matheson, 2001), or insert discourse markers (Kim et al., 2000). To the extent of this author’s knowledge, there has been no previous work in ranking questions for a tutorial dialogue context, nor has there been analysis into the role of dialogue act features for learning differences in tutoring style between experienced tutors.

7.2 Data Collection

7.2.1 MyST Logfiles and Transcripts

These experiments make use of the DISCUSS annotated MyST transcripts collected in the Wizard-of-Oz (WOZ) condition described in Chapter 5. In total 119 WOZ dialogues covering 10 units in magnetism and electricity and 2 in measurement and standards are used for building models and computing corpus statistics.

7.2.2 Question Authoring

Though a question generation system would provide the most deterministic way to control question variation, developing a fully-functional system free of grammar and fluency errors is beyond the scope of this thesis. While such a system is part of the long-term vision for this work, the experiments in this chapter make use of manually authored questions. This restriction shifts the
T: Tell me more about what is happening with the electricity in a complete circuit.
S: Well the battery sends all the electricity in a circuit to the motor so the motor starts to go.

<table>
<thead>
<tr>
<th>Candidate Question</th>
<th>Frame</th>
<th>Element</th>
<th>DISCUSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 Roll over the switch and then in your own words, tell me again what a complete or closed circuit is all about.</td>
<td>Same</td>
<td>Same</td>
<td>Direct/Task/Visual</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ask/Describe/Configuration</td>
</tr>
<tr>
<td>Q2 How is this circuit setup? Is it open or closed?</td>
<td>Same</td>
<td>Same</td>
<td>Ask/Select/Configuration</td>
</tr>
<tr>
<td>Q3 To summarize, a closed circuit allows the electricity to flow and the motor to spin. Now in this circuit, we have a new component. The switch. What is the switch all about?</td>
<td>Diff</td>
<td>Diff</td>
<td>Assert/Recap/Proposition</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Direct/Task/Visual</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ask/Describe/Function</td>
</tr>
<tr>
<td>Q4 You said something about the motor spinning in a complete circuit. Tell me more about that.</td>
<td>Same</td>
<td>Same</td>
<td>Revoice/None/None</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ask/Elaborate/CausalRelation</td>
</tr>
</tbody>
</table>

Table 7.1: Example dialogue context snippet and a collection of candidate questions. The frame, element, and DISCUSS columns show how the questions vary from one another.

focus away from issues of grammaticality and well-formedness and keeps the findings centered on learning to identify the factors driving selection of context appropriate questions in tutoring.

To maintain consistency in authoring and to avoid repeated efforts, a single author was used to write all of the candidate questions. The author, a linguist by training, was selected for his understanding of the processes driving variation in language. Consequently, initial training focused more on introductions to the FOSS curriculum and the QtA style of dialogue used in MyST tutoring. Next he was taught the DISCUSS taxonomy with a focus on varying questions lexically, syntactically, and semantically. Although he was free to author any question he found appropriate, the guidelines primarily emphasized authoring by making permutations aligned with DISCUSS dimensions while also permitting the author to incorporate changes in wording, learning-goal content, and tutoring tactics. For example, he was taught to consider how QtA moves such as Revoicing, Marking, or Recapping could alter otherwise similar questions. He also was asked to think about whether to continue the current line of questioning or to move on to a different topic. To minimize the risk of
rater bias, the author was explicitly told to avoid using positive feedback expressions such as “Good job!” or “Great!”. Table 7.1 illustrates how the combination of DISCUSS labels, QtA tactics, and dialogue context can drive the question generation process.

To simulate the conditions available to both the human WOZ and computer MyST tutors, the author was presented with the entire dialogue history preceding the decision point, the current dialogue frame (learning goal), and any visuals that may be on-screen. A sample screen shot of the web application used to collect questions is shown in Figure 7.1. Question authoring contexts were manually selected to capture points where students provided responses to tutor questions. This eliminated the need to account for other dialogue behavior such as greetings, closings, or meta-behavior, and kept candidate prompts in the form of follow-up style questions. Because these question authoring contexts came from actual tutorial dialogues, the original turn provided by the tutor were also extracted. Turns that did not contain questions related to the lesson content were filtered out. The corpus for these experiments has 205 question authoring contexts comprised of 1025 manually authored questions and 131 questions extracted from the original transcript yielding 1156 questions in total.
Given the dialogue context next, please author follow-up questions.

1. You said that the weight on one side was too heavy. How do you think the weight of the washers relates to the strength of the magnets?

2. You mentioned magnetic force. How do you think the magnetic force relates to the weight of the washers?

3. When the weight of the washers is too great, the magnets will be pulled apart. What's up with that?

4. What do you think would happen if there were no magnets at one end of the balance? Do you think that it would take more or less washers to tip the balance?

5. You mentioned something about magnetic force. What does the magnetic force have to do with the weight of the washers at the other end of the balance?

Figure 7.1: Question Authoring Tool Screenshot: The top (green) section shows the learning goals for the session. The middle (blue) section shows the dialogue history up to the question authoring point. The bottom section consists of textboxes for accepting the author’s questions.

7.2.3 Ratings Collection

Four tutors who had previously served as project tutors and wizards were recruited to provide ratings for the questions. These raters were presented with much of the same information used
during question authoring. The interface included the entire dialogue history preceding the question decision point and a list of up to 6 candidate questions (5 manually authored, 1 taken from the original transcript if applicable). To give a more complete tutoring context, raters also had access to the lessons’ learning goals and the interactive visuals used by MyST. A screenshot of the ratings collection application is shown in Figure 7.2

Previous studies in rating questions (Becker et al., 2009) have found poor inter-rater agreement when rating questions in isolation. To decrease the task’s difficulty raters were instead asked to score all candidate questions simultaneously. To avoid biasing rater decisions, no specific criteria for question quality were defined. Instead the raters were instructed to consider the question’s role in assisting student understanding of the learning goals and to think about factors such as tutorial pacing, context appropriateness, and content. Scores were collected using an ordinal 10-point scale ranging from 1 (lowest/worst) to 10 (highest/best).

Each set of questions was rated by at least three tutors, and rater assignments were selected to ensure raters never score questions from sessions they tutored themselves. In total ratings were collected for 1156 questions representing a total of 205 question contexts distributed across 30 transcripts.

7.2.4 Rater Agreement

Because the criteria for judging question quality are highly subjective, a key challenge in this work centers on understanding to what degree the tutors agree with one another. Since the final task is to rank questions and not to score questions, agreement is assessed in terms of rank-agreement rather than proximity of score values. To do this each tutors’ scores with a given context are converted into a rank-ordered list. Kendall’s-Tau (τ) rank correlation coefficient is then used to compute inter-rater agreement in ranking. This measure is a non-parametric statistic that quantifies the similarity in orderings of data, and it is closely tied to AUC, the area under the receiver operating characteristics (ROC) curve. Though Kendall’s-τ can vary from -1 to 1, its value is highly task dependent, and it is typically lower when the range of possible choices is narrow as it is in this task.
Figure 7.2: Question Rating Form Screenshot: The top (green) section shows the learning goals for the session. The middle (blue) section shows the dialogue history up to the question authoring point. The bottom section consists of textboxes for accepting the author’s questions.

To get a single score $\tau$ values are averaged across all sets of questions (contexts) and all pairs of raters.

The mean value for all pairs of raters and contexts is $\tau = 0.1478$, and the rater versus rater agreement statistics are shown in Table 7.2. Additionally, a given pair of raters agreed on the top-rated question 33% of the time. While inter-rater agreement is fairly modest, there is wide variation dependent on the pair of tutors. This suggests that despite their common training and experience, the raters may be using different criteria in rating.

To assess the tutors’ internal consistency, each tutor re-rated 60 sets of questions approximately two months after their first trial. Self-agreement Kendall’s-$\tau$ values were computed using the procedure described above. These statistics are listed in the bottom row of Table 7.2. In contrast
with the inter-rater agreement, self-agreement is much more consistent providing further evidence for a difference in criteria. Together self and inter-rater agreement help bound expected system performance in ranking.

<table>
<thead>
<tr>
<th></th>
<th>rater A</th>
<th>rater B</th>
<th>rater C</th>
<th>rater D</th>
</tr>
</thead>
<tbody>
<tr>
<td>rater A</td>
<td>X</td>
<td>0.2590</td>
<td>0.1418</td>
<td>0.0075</td>
</tr>
<tr>
<td>rater B</td>
<td>0.2590</td>
<td>X</td>
<td>0.1217</td>
<td>0.2370</td>
</tr>
<tr>
<td>rater C</td>
<td>0.1418</td>
<td>0.1217</td>
<td>X</td>
<td>0.0540</td>
</tr>
<tr>
<td>rater D</td>
<td>0.0075</td>
<td>0.2370</td>
<td>0.0540</td>
<td>X</td>
</tr>
<tr>
<td>mean</td>
<td>0.1361</td>
<td>0.2059</td>
<td>0.1058</td>
<td>0.0995</td>
</tr>
<tr>
<td>self</td>
<td>0.4802</td>
<td>0.4022</td>
<td>0.2327</td>
<td>0.3531</td>
</tr>
</tbody>
</table>

Table 7.2: Inter-rater rank agreement (Kendall’s-τ): The bottom row is the self-agreement for contexts they rated in two separate trials.

7.3 Automatic Ranking

Like with evaluating rater agreement, the goal of this study is to learn to predict which questions are more suitable for a given tutoring scenario and not necessarily to assign specific scores to questions. This task of question selection is framed as a ranking task embedded within a supervised machine learning framework. To create a gold standard for training and evaluation the collective ratings for a set of questions are first converted into a rank-ordered list. While the most straightforward way to make this conversion is to average the ratings for each item, this approach assumes all raters operate on the same scale. Furthermore, a single score does not account for how a question relates to other candidate questions. Instead the process for creating a single rank-order begins with tabulation of pairwise wins for all pairs of questions $q_i, q_j, (i \neq j)$ within a given dialogue context $C$. If $\text{rating}(q_i) > \text{rating}(q_j)$, questions $q_i$ receives a win. This is summed across all raters for a given context. The question(s) with the most wins has rank 1. Questions with an equal number of wins are considered tied and are given the average ranking of their ordinal positions. For example if two questions are tied for second place, they are each assigned a ranking of 2.5.

Using this rank-ordering a pairwise classifier is trained to learn a binary preference function
(Cohen et al., 1998) that determines if one question has a better rank than another. To do this, a vector of features $\phi_i$ is constructed for each question $q_i$ within a dialogue context $C$. For each pair of questions $q_i$ and $q_j$, a new vector is created using the difference of features: $\Phi(q_i, q_j, C) = \phi_i - \phi_j$. For training, if $\text{rank}(q_i) < \text{rank}(q_j)$, the classification is positive otherwise it is negative. To account for the possibility of ties, and to make the difference measure appear symmetric, both combinations $(q_i, q_j)$ and $(q_j, q_i)$ are trained. During decoding, the trained classifier is run on all pairs and wins are tabulated using the approach described above.

For these experiments pairwise classifiers are trained using Mallet’s Maximum Entropy (McCallum, 2002) and SVMLight’s Support Vector Machines models (Joachims, 1999). SVMRank (Joachims, 1999) is also used, which performs the same maximum margin separation as SVMLight, but effectively uses Kendall’s-$\tau$ as a loss function to optimize for rank ordering. The SVMRank classifier is run with a linear kernel and model parameters of $c = 2.0$ and $\epsilon = 0.0156$. For MaxEnt, model parameters are set to Mallet’s defaults. Training and evaluation are carried out using 10-fold cross validation (3 transcripts per fold, approximately 7 dialogue contexts per transcript). Folds are partitioned by FOSS unit, to ensure training and evaluation are on different lessons. To explore the impact of DISCUSS representations on this question ranking task, models are trained and evaluated by incrementally adding additional information extracted from the DISCUSS annotation.

To assess the systems’ abilities in replicating question ranking behavior, models are trained and evaluated in the following conditions:

<table>
<thead>
<tr>
<th>Training / Evaluation</th>
<th>Individual Rankings</th>
<th>Combined Rankings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Rankings</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Combined Rankings</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Throughout the rest of this chapter, the term **General Model** will refer to models trained on the rankings combined from multiple raters, and the term **Individual Model** will refer to those trained on rankings from a single rater.
Feature design for this task was motivated by the desire to capture the factors that may play a role in the tutor’s decision making process during question selection. When rating, scorers may consider factors such as the question’s surface form, use of lesson-specific keywords, contextual relevance, etc. These features effectively serve as hypotheses about what drives the question asking process. Table 7.3 shows a complete list of the features, the subsections below detail the motivations and intuitions behind these factors.

### 7.4.1 Surface Form Features

When presented with a list of questions, a rater likely bases the decision on his or her initial reaction to the questions’ wording. In some cases, wording may supercede any other decisions regarding educational value or dialogue cohesiveness. Question verbosity is captured by the **number of words in the question** feature. Analysis of rater comments also suggested that preferences are often tied to the question’s form and structure. A rough measure of form comes from the **Wh-word** features to mark the presence of the following question words: who, what, why, where, when, which,
and how. Additionally the **bag-of-part-of-speech-tags (POS)** features provide another aspect of the question’s structure.

### 7.4.2 Lexical Similarity Features

Past work (Ward et al., 2011a) has shown that entrainment, the process of automatic alignment between dialogue partners, is a useful predictor of learning and is a key factor in facilitating a successful conversation. For question selection, one natural hypothesis follows from the notion that successful tutors ask questions that display some degree of semantic entrainment with student utterances. In MyST-based tutoring, dialogue actions are driven by the goal of eliciting student responses that address the learning goals for the lesson. Consequently, choosing an appropriate question may depend on how closely student responses align with the learning goals. To model both entrainment and lexical similarity the following set of features are extracted: unigram and bigram word-overlap, lemma-overlap, and part-of-speech (POS) tag overlap. These overlap values are computed between the pairs of text listed below.

- The candidate question and the student’s last utterance
- The candidate question and the last tutor’s utterance
- The candidate question and the text of the current learning goal
- The candidate question and the text of the other learning goals

Example learning goals for a lesson on circuits are provided in Table 7.4. The current learning goal is simply the learning goal in focus at the point of question asking according to the MyST logfile. Other learning goals are all other goals for the lesson. Using the example from the table, if goal 2 is the current learning goal, then goals 1 and 3 are the other goals.

### 7.4.3 DISCUSS Features

The lexical and surface form features provide some cues about the content of the question, but they do not account for the action or intent in tutoring. The DISCUSS annotation bridges the
Goal 1: *Wires carry electricity and can connect components*
Goal 2: *Bulb receives electricity and transforms electricity into heat*
Goal 3: *A circuit provides a pathway for energy to flow*

<table>
<thead>
<tr>
<th>Table 7.4: Example Learning Goals</th>
</tr>
</thead>
</table>

question’s semantics and pragmatics and focuses on what differentiates one question from another. Basic DISCUSS features include bags of Dialogue Acts (DA), Rhetorical Forms (RF), and Predicate types (PT) found in the question’s DISCUSS annotation. The question’s dialogue cohesiveness is captured using binary features indicating whether or not the question’s RF and PT match those found in the previous student and tutor turns.

### 7.4.4 Contextualized DISCUSS Features

In tutoring, follow-up questions are licensed by the questions that precede them. For example a tutor may be less likely to ask how an object functions until after the object has first been identified by the student. Along a different dimension, a tutor’s line of questioning may change to match a student’s understanding of the material. Struggling students may require additional opportunities to explain themselves, while advanced students may benefit more from a more rapid pace of instruction.

The conditional relevance of questions is modeled by computing dialogue act transition probabilities from the corpus of DISCUSS annotated tutorial dialogues. Although DISCUSS allows multiple tags per dialogue turn, probability calculations are simplified by treating each DISCUSS tuple as a separate event, and tallying all pairs of turn-turn labels. A DISCUSS tuple consists of a Dialogue Act (DA), Rhetorical Form (RF), and Predicate Type (PT), and different subsets of the tuple are used to compute the transition probabilities listed in Equations (7.1) to (7.3). All probabilities are computed using Laplace smoothing. When extracting features, the log of the probabilities are summed for each DISCUSS label present in the question.

MyST models dialogue as a sequence of semantic frames which correspond to specific learning
goals. For natural language understanding, MyST uses Phoenix semantic grammars (Ward, 1994) to identify which elements within these frames have been filled. The conditional probability of a DISCUSS label given the percentage of elements filled in the current dialogue frame (Equation (7.4)) provides an account of student progress during question asking. This progress percentage is discretized into quartiles of 0-25%, 25-50%, 50-75%, and 75-100%.

\[
p(DA, RF, PT_{question}|DA, RF, PT_{stud. turn})
\]
\[
p(DA, RF_{question}|DA, RF_{student turn})
\]
\[
p(PT_{question}|PT_{student turn})
\]
\[
p(DA, RF, PT_{ques.|\% elements filled})
\]

### 7.5 Evaluation

System performance is evaluated using two measures commonly utilized in information retrieval: the Mean Kendall’s-$\tau$ measure described in Section 7.2.4 and Mean Reciprocal Rank (MRR). As was done for inter-rater agreement, the Tau-b variant of Kendall’s-$\tau$ is used in lieu of the standard Kendall’s-$\tau$ for its ability to factor in ties into the overall score. MRR is the average of the multiplicative inverse of the rank of the highest ranking question across all contexts. The higher the MRR the better, with 1 being the best possible score. For a given question asking context, MRR is computed by taking the inverse of the average system-assigned rank for all questions ranked first in the gold standard. The gold standard ranking in this comparison is the combination of all raters’ scores for the general model and is the scores for a specific rater in the individual models.

### 7.6 Results and Discussion

#### 7.6.1 General Model: Results

Several models were trained and evaluated to investigate how different feature classes influence overall performance in ranking. The results for these experiments are listed in Table 7.5. Because
MaxEnt and $SVM^{Light}$ exhibited comparable performance, only the results for MaxEnt and $SVM^{Rank}$ models are reported. In addition to MRR and Kendall’s-$\tau$, the pairwise classification concordances and discordances counts are listed to give the reader another sense of the accuracy associated with rank agreement.

**Random Baseline:** On average, assigning random ranks will yield mean $\tau$=0 and MRR=0.408.

**Baseline System:** The baseline system used all of the surface form and lexical similarity features described above. This set of features achieves the highest rank agreement ($\tau = 0.105$) using maximum entropy and the highest MRR (0.473) with $SVM^{Rank}$. This improvement over the random baseline suggests there is a correlation between a question’s ranking and its surface form.

**DISCUSS System:** Table 7.5 shows system performance steadily improves as additional DISCUSS features are included in the model. When using DISCUSS features, removing the part-of-speech features gives an additional bump in performance suggesting that there is an overlap in information between DISCUSS representations and POS tags. Finally, adding contextualized DISCUSS features pushes the ranking models to their highest level of agreement with $\tau = 0.211$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Mean Kendall’s-$\tau$</th>
<th>Num. Concord.</th>
<th>Num. Discord.</th>
<th>Pairwise MRR</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaxEnt</td>
<td>CONTEXT+DA+PT+MATCH+POS-</td>
<td>0.211</td>
<td>1560</td>
<td>974</td>
<td>0.616</td>
<td>0.516</td>
</tr>
<tr>
<td>SVM$^{Rank}$</td>
<td>CONTEXT+DA+PT+MATCH+POS-</td>
<td>0.190</td>
<td>1725</td>
<td>1154</td>
<td>0.599</td>
<td>0.555</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>CONTEXT+DA+RF+PT+MATCH+POS-</td>
<td>0.185</td>
<td>1529</td>
<td>1014</td>
<td>0.601</td>
<td>0.512</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>DA+RF+PT+MATCH+POS-</td>
<td>0.179</td>
<td>1510</td>
<td>1009</td>
<td>0.599</td>
<td>0.503</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>DA+RF+PT+MATCH+</td>
<td>0.163</td>
<td>1506</td>
<td>1044</td>
<td>0.591</td>
<td>0.485</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>DA+RF+PT+</td>
<td>0.147</td>
<td>1500</td>
<td>1075</td>
<td>0.583</td>
<td>0.480</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>DA+RF+</td>
<td>0.130</td>
<td>1458</td>
<td>1082</td>
<td>0.574</td>
<td>0.476</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>DA+RF</td>
<td>0.120</td>
<td>1417</td>
<td>1076</td>
<td>0.568</td>
<td>0.458</td>
</tr>
<tr>
<td>SVM$^{Rank}$</td>
<td>Baseline</td>
<td>0.108</td>
<td>1601</td>
<td>1278</td>
<td>0.556</td>
<td>0.473</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>Baseline</td>
<td>0.105</td>
<td>1410</td>
<td>1115</td>
<td>0.558</td>
<td>0.448</td>
</tr>
</tbody>
</table>

Table 7.5: Question ranking system scores by feature set and machine learning model: Presence or absence of specific features is denoted with a ‘+’ or ‘-’ otherwise the label refers to a set of features. The Baseline features consist of the Surface Form and Lexical Similarity features described in Sections 7.4.1 and 7.4.2. POS are the bag-of-POS surface form features. DA, RF, and PT refer to the DISCUSS presence features for the Dialogue Act, Rhetorical Form, and Predicate Type dimensions described in Section 7.4.3. MATCH refers specifically to the RF and PT match features. CONTEXT refers to the Contextualized DISCUSS features described in Section 7.4.4. The best scores for each column appear in boldface.
using MaxEnt and $\text{MRR}=0.555$ using $\text{SVM}^{\text{Rank}}$. Inspection of the MRR values shows that without taking into account the possibility of ties the baseline system selects the top-ranked question in 44 out of 205 contexts (21.4%). While the system with the best MRR score correctly chooses the
top-ranked question in 71 out of 205 contexts (34.6%) – a rate comparable to how often a pair of raters agreed on the number-one item (33.4%).

Application of the Wilcoxon signed-rank test shows the DISCUSS system exhibits statistically significant improvement over the baseline system in its distribution of Kendall’s-τ values ($n = 205, z = 7350, p < 0.001$) and distribution of reciprocal ranks ($n = 205, z = 3739, p < 0.001$). Figures 7.3 and 7.4 give visual confirmation of this improvement, and highlight the overall reduction in negative τ values as well as the greater-than-50% increase in likelihood of selecting the best question first.

To get another perspective on system performance, rankings from the human raters were scored against the gold standard rankings from the subset of questions used for assessing internal agreement. This yielded a mean τ between 0.2589 and 0.3619. If ratings are removed so that the gold standard does not include the rater under evaluation, tutor performance drops to a range of 0.1523 to 0.2432, which is roughly centered around the agreement exhibited by the best-performing system.

Looking at the impact of learning algorithms, one can see that $SVM^{Rank}$ tends to perform better on the MRR metric while the pairwise maximum entropy models yield higher τ-values. One possible explanation for this discrepancy may stem from the ranking algorithms’ different treatment of ties. The pairwise model permits ties, whereas the scores produced by $SVM^{Rank}$ produce a strict order. Without ties, it is difficult to exactly match the raters’ orderings which had numerous ties, which can in turn produce an overall higher number of concordances and discordances than the pairwise classification model.

The general model’s major limitation most likely stems from inter-rater agreement in ranking. While the Kendall’s-τ values show agreement above chance, the scores may suggest that there are bi-modal distributions in assigning scores. For a given question one class of tutors may score it very highly whereas another may score it on the low end of the scale. Considering that all four tutors have gone through identical training with respect to QtA and MyST tutoring, this variance is less a commentary on task difficulty and more a reflection on differences in preferences and
tutoring styles. As such, the performance of the general model can be taken as evidence that the DISCUSS augmented feature space provides information rich enough to capture the factors involved in choosing follow-up questions.

### 7.6.2 Individual Models

For the sake of interpretability, only Maximum Entropy Models are used for training and evaluation on individual tutors’ rankings. Like in the general model condition, several models were trained using different subsets of the feature categories. Table 7.6 lists the Kendall’s-$\tau$ rank order agreement for models trained on individual tutors as well as the combined model. Applying the Wilcoxon-signed rank test to the distribution of Kendall’s-$\tau$ values (i.e. per dialogue context agreement coefficients) shows a statistically significant improvement ($p < 0.01, 148 \leq n \leq 161$) between the baseline and top-performing models for all raters. This suggests that features extracted from DISCUSS provides additional information not available in the surface form and lexical similarity features. However, performance is not strictly tied to the number of DISCUSS-based features. Unlike the models trained on average rankings, models trained to replicate an individual rater’s rankings may require only a subset of the total features. For example, the best model for Rater C used only the dialogue act and baseline features, whereas Rater D showed improvement when adding the more complex contextual features. These differences in performance roughly outline what level of linguistic detail underlies a rater’s preference for one question over another.

Comparing the results from Table 7.6 with the inter and intra-rater agreement values from Table 7.2 shows that the models best able to replicate individuals’ rankings are those trained on data from raters with the highest self-agreement. This suggests that data collection should be improved to limit variation in judgment. One potential way to improve rater reliability would be to back away from having raters simultaneously scoring all questions and instead present them with paired comparisons.
Table 7.6: System mean Kendall’s-τ rank-order agreement scores by model and rater. Model training and evaluation is conducted per rater, or in the case of All, a combination of the four raters. The General Model row shows agreement between output from a system trained on the combination of raters (the best model in the 'All' column) and the gold standard rankings from individual raters. Presence or absence of features is denoted with a ‘+’ or ‘-’. The Baseline features consist of the Surface Form and Lexical Similarity features.

7.6.3 Feature Analysis:

To get a qualitative perspective for the individual models, a brief description of each rater’s tutoring style was collected from the lead tutor on the MyST project. Specifically she was asked to provide a one-line summary of their approach to tutoring based on her observations of their tutoring during the MyST-WOZ study. She offered:

- “Rater A focuses more on the student than the lesson.”

- “Rater B focuses on the lesson objectives.”

- “Rater C tries to get the student to relate to what they see or do.”

- “Rater D likes to add more to the lesson than what was done in class.”

Looking at the models in light of these comments, one sees the feature weights reflect these differences in tutoring philosophies. Rater A’s model was the only one to give a negative weight to the Assert DA feature, which may stem from a desire to elicit speech instead of lecture. Rater B’s emphasis on learning goals manifests itself with larger weights for the lexical overlap features than the other rater
models. Rater C’s emphasis on visuals results in PT features weighted towards **Observation** over **Function** or **Process**. Rater D’s desire to create a new experience yields a DA **Metastatement** weight that is twice that found in the other raters’ models. Additionally, rater D had the largest weight for the contextual probability features.

Table 7.7 shows the feature category distributions for the 20 (15%) most influential features (those with the largest weight magnitudes), and shows a wide variance from rater to rater. Figure 7.5 shows a graphical representation of this information. Inspecting these data more closely, provides more evidence that the models are learning the distinctions between the tutors’ preferences. For example rater B’s emphasis on the lesson objectives parallels the higher weights given to the baseline features, which are largely composed of lexical overlap between the utterances and the learning goals. Conversely, the largest percentage of rater C’s top features are found in the predicate type category, which follows closely with rater C’s emphasis on visuals. Only the predicate type features convey this information.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>DA</th>
<th>RF</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater A</td>
<td>0.163</td>
<td>0.312</td>
<td>0.245</td>
<td>0.281</td>
</tr>
<tr>
<td>Rater B</td>
<td>0.557</td>
<td>0.123</td>
<td>0.134</td>
<td>0.187</td>
</tr>
<tr>
<td>Rater C</td>
<td>0.275</td>
<td>0.200</td>
<td>0.195</td>
<td>0.330</td>
</tr>
<tr>
<td>Rater D</td>
<td>0.581</td>
<td>0.114</td>
<td>0.101</td>
<td>0.204</td>
</tr>
<tr>
<td>All</td>
<td>0.374</td>
<td>0.151</td>
<td>0.139</td>
<td>0.336</td>
</tr>
</tbody>
</table>

Table 7.7: Distribution of the 20 most influential rater model features by coarse category.

<table>
<thead>
<tr>
<th></th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater C</th>
<th>Rater D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater A</td>
<td>1.000</td>
<td>0.526</td>
<td>0.167</td>
<td>0.163</td>
</tr>
<tr>
<td>Rater B</td>
<td>0.526</td>
<td>1.000</td>
<td>0.106</td>
<td>0.250</td>
</tr>
<tr>
<td>Rater C</td>
<td>0.167</td>
<td>0.106</td>
<td>1.000</td>
<td>0.184</td>
</tr>
<tr>
<td>Rater D</td>
<td>0.163</td>
<td>0.250</td>
<td>0.184</td>
<td>1.000</td>
</tr>
<tr>
<td>mean</td>
<td>0.464</td>
<td>0.470</td>
<td>0.364</td>
<td>0.399</td>
</tr>
</tbody>
</table>

Table 7.8: Rater Model Cosine Similarities: similarities are computed using the feature weights for each pair of rater models.
While the distribution of top-20 feature categories gives some confidence that the behavior produced by the learned models corresponds to their real-world counterpart, they are too coarse to understand how they align and vary from rater to rater. Consider raters A and D. Although each of these models have very similar top-20 distributions, they were not actually the most closely aligned in inter-rater agreement. Cosine similarity provides another metric for quantifying the relationship between models. Specifically, the cosine similarity measures the similarity between two vectors by measuring the cosine of the angle between them. For a pair of raters $i$ and $j$, the cosine similarity is computed using the vectors of feature weights ($\beta_i, \beta_j$) learned during model training (Equation (7.5)). The cross-tabulation for all pairs of rater models is listed in Table 7.8. Comparing this table of cosine similarity scores to the Kendall’s-$\tau$ values listed in Table 7.2 reveals...
parallels between model agreement and inter-annotator agreement. In three out of four instances, the highest cosine similarity values for each model correspond to the highest Kendall’s-τ values for each rater, suggesting that the feature space used for this question ranking task accounts for many of the factors used by tutors during questioning.

\[
similarity = \cos(\theta) = \frac{\beta_i \cdot \beta_j}{||\beta_i|| ||\beta_j||} \tag{7.5}
\]

### 7.6.4 Error Analysis:

Cross-referencing rater feedback with analyses of contexts with low system-rater agreement helped to identify three categories of errors: 1) question authoring errors, 2) DISCUSS annotation errors and 3) model deficiency errors. Example question authoring mistakes include referencing an interactive visual when a static one was on-screen, and writing questions which were too wordy or used incorrect terminology. Unlike the raters, the system was unable to key in on these mistakes. While better quality control would help to reduce many of these errors, for future work in fully-automatic question generation, language model or vocabulary features may help to give a better account of the suitability or age-appropriateness of a question’s surface form. In instances with incorrect DISCUSS annotation, having a correct label would have likely yielded better classification accuracy and consequently ranking agreement. In one context all of the candidate questions had nearly identical DISCUSS representation; however the question that should have been ranked more highly by the system had a wrongly labeled predicate type. Although student learning goal completion is factored into the DISCUSS context probability features, a large proportion of errors coincided with rater comments about student understanding and misconceptions. This suggests that additional features that capture student correctness such as answer-question entailment could benefit system performance.
7.7 Conclusions

This chapter has introduced a framework for learning and evaluating models for ranking and selecting questions for a given point in a tutorial dialogue. Furthermore these experiments show that it is feasible to learn this behavior by coupling predefined questions with ratings from trained tutors. Adding features extracted from the dialogue context and DISCUSS dialogue act annotation to baseline surface form and lexical similarity features improves system performance in ranking to a level in accordance with experienced human tutors. These results illustrate how question asking depends not only on the form of the question but also on the underlying dialogue action, function and content. Lastly, this framework provides a natural starting point to explore the use of the DISCUSS dialogue move representations for automatic question generation.
A central issue in tutoring research lies in identifying and understanding the strategies that lead to long-term retention of the material. While the efficacy of human and computer tutoring is well established, the mechanisms that make for a successful tutorial dialogue are still not well understood. This poses a challenge for implementation of intelligent tutoring systems as there are numerous authoring and design decisions which hinge not only on the subject matter but on underlying pedagogical philosophies.

The work in this chapter builds on past studies that explore how features of tutorial dialogue correlate with learning by adding features extracted from Phoenix semantic parses and DISCUSS tags. The findings help to confirm existing notions and discover new insights about the nature of dialogue in MyST and other intelligent tutoring systems. Analysis of dialogue act correlations found that QtA-style elements such as Feedback and Mark have positive correlations with learning, while excessive instruction has a negative correlation. From a processing perspective, these results demonstrate that automatic move classifiers can provide reliable signals for deeper analysis of tutorial dialogue.

8.1 Background and Related Works

In the past decade, there has been increased attention on using empirical measures of student performance to motivate design of dialogue-based intelligent tutoring systems. Shallow measures such as the percentage of words in the dialogue uttered by the student and average length of a
student turn have been shown to have a positive correlation with learning gains (Core et al., 2003; Katz et al., 2003; Rosé et al., 2003). Though these findings support hypotheses regarding the role of student initiative in dialogue (Core et al., 2003) and the importance of self-explanation (Chi et al., 2001, 1994), on their own they neither explain outcomes nor inform design of tutoring systems.

Other studies have relied on coding transcripts with tutoring acts and dialogue acts (Pilkington, 1999; Buckley and Wolska, 2008) to gain deeper insight into the tutoring process. Dialogue coding has been applied to investigate common collaborative patterns in tutoring (Graesser et al., 1995) or to explore the relation between student questioning and learning (Graesser and Person, 1994). Jackson et al. (2004) used similar coding schemes to learn that student learning is positively correlated with tutoring acts that force the student to provide the majority of the answer and is negatively correlated in cases where the tutor gives the answer away. Similarly Litman and Forbes-Riley (2006) and Litman et al. (2009) used both unigram and bigram analysis of dialogue acts to “show that student introductions of a new concept into the dialogue positively correlates with learning, but student attempts at deeper reasoning (particularly when incorrect), and the human tutors’ attempts to direct the dialogue, both negatively correlate with learning”. Lastly, Boyer (2010) utilized Hidden Markov Models to discover latent dialogue modes which correlate with learning.

This work extends the idea of correlating dialogue act features with measured gains to feature spaces composed of rich semantic and pragmatic representations of dialogue. The finer-grained details afforded by DISCUSS moves and Phoenix semantic parses provide additional insight into the composition of the questions and responses made by tutors and students respectively. Furthermore this is one of the few works to report dialogue act correlations derived from automatically labeled dialogues.

8.2 Student Outcomes Data

Student outcomes data used for learning correlations were collected for one-hundred and two (102) students in 14 classrooms in 6 schools using the MyST standalone condition. More details about this data collection are provided in Section 2.6.
8.2.1 Residual Learning Gains

Perhaps the most predictive measure of learning gain (i.e. the difference between post-test and pre-test scores) is the pre-test itself. A student who scores well in the pre-test likely knows the material already and does not have much room for additional gains. Conversely, for a student who scores poorly on the pre-test even an improvement towards the average will bring a large learning gain. Because of this tight coupling between pre-test and post-test scores, raw learning gain is not a useful measure for analyzing the factors contributing to learning. Instead the analyses below rely on a measure referred to as the residual learning gain (Willett, 1988).

In statistics the residual of a sample is the difference between the sample and the estimated (expected) value. When measuring change, the residual is used to remove correlations between pre- and post-conditions; thus residual learning gain is a metric that decouples the post-test score from its corresponding pre-test score. It is computed by first using a function to compute an estimate of the post-test score given the pre-test score (Equation (8.1)). The beta ($\beta$) weights for the estimation function are obtained using least squares linear regression.

$$\hat{\text{score}}_{\text{post}} = \beta_1 \times \text{score}_{\text{pre}} + \beta_0$$  \hspace{1cm} (8.1)

Residual learning gain is then the difference between the actual post-test and the expected post-test score (Equation (8.2)).

$$\text{gain}_{\text{residual}} = \text{score}_{\text{post}} - \hat{\text{score}}_{\text{post}}$$  \hspace{1cm} (8.2)

Because the magnitude of gains can vary across different FOSS modules and investigations, residual gain is normalized to a percentage by dividing by the maximum possible score for an investigation. Unless specified otherwise, the terms learning gains and residual learning gains will refer to this normalized residual learning gain for the remainder of this chapter.
8.2.2 Learning Outcomes Granularity

For assessing the effect of MyST tutoring, the module-level ASK assessment is sufficient for measuring student progress. However, this granularity poses a problem for dialogue-based analysis as one module may have up to 16 tutoring dialogues associated with it. To get a true coupling between individual tutoring sessions and measured learning gains would require a pre-test before and a post-test after each tutoring session. The coordination and organization to orchestrate such an effort across multiple schools and multiple lessons was well beyond the scope of the MyST study.

To make the most use of the existing transcripts and data necessitated a different approach. Instead of requiring exams scores for every transcript, the module-level ASK scores were decomposed into investigation level scores using grading rubrics included by FOSS. These rubrics specified which concepts and materials were covered by each of the questions in the assessment tests, which in turn were used to compute residual learning gains at the FOSS investigation level. This also means that correlations and analysis between dialogue and student outcomes occur with a correspondence of up to four (4) tutorial dialogues per score.

8.3 DISCUSS Annotation

The volume and diversity of transcripts from the MyST assessment study made it impractical to undertake another project to manually annotate another corpus of transcripts with DISCUSS moves. Instead two automated flows were developed to annotate dialogue turns from both tutor and students.

8.3.1 Tutor Turns

Unlike the more free form Wizard-of-Oz condition, tutor utterances in the standalone MyST environment are restricted to those specified in the Phoenix dialogue manager configuration files. Consequently, the cost of manual DISCUSS annotation can be amortized across all logs with a smaller upfront cost. Instead of hand-annotating each individual utterance, the predefined turns
from the system configuration files were labeled once, and subsequent labeling of the tutor utterances from system-generated dialogues were performed automatically via lookup tables.

To create this flow, the author of this thesis manually tagged the utterances within the MyST “speak.txt” files with DISCUSS tuples. A sample snippet of the resulting “speak.discuss” file is shown in Figure 8.1. In total 3251 unique utterances were annotated with DISCUSS tuples. These utterances come from 38 speak files from 12 different investigations spanning three FOSS modules (Magnetism and Electricity, Measurement, and Variables).

![Figure 8.1: Example speak.discuss File](image)

8.3.2 Student Turns

Student turns were automatically labeled using the DISCUSS classifiers described in Chapter 6. Models were trained using the merged DISCUSS labels and the parameters listed in Table 6.3. Because the DISCUSS student turn annotator uses a collection of binary classifiers, it produces a bag-of-labels for an utterance in lieu of the DISCUSS tuples obtained from manual annotation.
8.4 Data Cleansing

Though the learning gains study tried to control for time on task, not all students participated in the same number of tutoring sessions between their pre and post tests. To mitigate the effects of time on task, the data set was further filtered to only those students who had participated in 3 (the median) or more tutoring sessions per investigation, reducing the data set to 168 instances, where an instance is the collection of dialogues for a student at a granularity of FOSS investigation.

8.5 Correlations with Features

8.5.1 Basic Features

MyST logfiles include basic, low-level information such as speaker timestamps, speech recognition output, and transcribed speech. With minimal processing, these data are transformed into features based on elapsed time, word counts, or turn counts. Statistics based on these counts are coarse measures of dialogue initiative and serve as high-level hypotheses on the role of student and tutor participation in tutoring. For example, one would expect that getting a student to speak more should improve his likelihood of learning. Conversely if the tutor leads too much of the conversation the student may not benefit as much from the dialogue. Table 8.1 lists the correlations between the basic features and learning gains. Descriptions of the basic features can be found in Appendix C.

Though slight, these correlations align with intuitions about student engagement and tutorial dialogue. Features exhibiting a positive correlation with learning include the number of words per student turn and the proportion of the total dialogue time spoken by the student – lending additional support to past findings that suggest increased speech from the student leads to increased learning gains. In contrast to the student turn statistics, tutor features such as average number of words per tutor turn showed a negative correlation with learning. One explanation is that increased lecturing can lead a student to tune out, alternatively increased initiative from the tutor may also
Table 8.1: Basic Dialogue Features Correlations: $R$ is Pearson’s correlation coefficient between the feature and the residual learning gain, and $p$ is the p-value for testing non-correlation. Rows are ordered by descending $R$ values.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>$R$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE_TURN_TIME_STUDENT (seconds)</td>
<td>12.07</td>
<td>4.26</td>
<td>0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>AVERAGE_WORDS_PER_TURN STUDENT</td>
<td>18.39</td>
<td>7.77</td>
<td>0.09</td>
<td>0.30</td>
</tr>
<tr>
<td>PERCENT_TIME_STUDENT</td>
<td>0.55</td>
<td>0.09</td>
<td>0.09</td>
<td>0.28</td>
</tr>
<tr>
<td>PERCENT_WORDS_STUDENT</td>
<td>0.40</td>
<td>0.10</td>
<td>0.09</td>
<td>0.28</td>
</tr>
<tr>
<td>AVERAGE_TURN_TIME_TUTOR (seconds)</td>
<td>6.93</td>
<td>1.26</td>
<td>-0.02</td>
<td>0.79</td>
</tr>
<tr>
<td>PERCENT_TURNS_STUDENT</td>
<td>0.41</td>
<td>0.02</td>
<td>-0.11</td>
<td>0.19</td>
</tr>
<tr>
<td>AVERAGE_WORDS_PER_TURN TUTOR</td>
<td>17.65</td>
<td>1.72</td>
<td>-0.16</td>
<td>0.05</td>
</tr>
<tr>
<td>AVERAGE_ONSET_TIME_STUDENT (seconds)</td>
<td>3.57</td>
<td>2.12</td>
<td>-0.16</td>
<td>0.06</td>
</tr>
<tr>
<td>AVERAGE_ONSET_TIME_TUTOR (seconds)</td>
<td>3.05</td>
<td>3.63</td>
<td>-0.22</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Lastly, speaker onset times for both tutors and students correlated negatively with learning gains. In terms of magnitude these two features had the strongest correlation among all of the basic features. Intuitively this reflects a link between tutor-student responsiveness and overall learning. While short onset times likely signal speaker responsiveness longer delays may stem from issues with the system or student uncertainty. Regardless of the source of onset, an extended delay makes for a less realistic MyST experience.

Though these correlations align with expectations about tutor and student behavior, they provide only a coarse-level of insight into the dialogue sessions. The following sections use more detailed dialogue features to better characterize and understand what makes for successful MyST tutoring sessions.

8.5.2 Phoenix Dialogue Manager Features

In addition to time stamps and the dialogue text the Phoenix Dialogue Manager within MyST is a source for more detailed information about the system dialogue state. Within MyST, a dialogue is modeled as a series of conversations revolving around different semantic frames. Much of the work in authoring MyST dialogues centers on writing Phoenix task files to define the frames, slots.
(frame-elements) and **rules** that govern the flow of the conversation. An example task file is shown in Figure 8.2. Frames usually correspond directly to a FOSS learning goal for the unit and the slots represent the semantic decomposition of the frame. The dialogue manager’s goal is to elicit speech from the student that fills the slots. Each slot is associated with a set of prompts containing dialogue actions such as synthesizing text or presenting a visual. The dialogue manager will iterate through each of the prompts for a slot until either the student has filled the slot or until all of the prompts have been exhausted. Additionally, the values of slots can trigger rules which like slots are associated with one or more prompts.

Because there is no ground-truth for student correctness, knowledge of the system state throughout the dialogue can serve as a proxy for assessing the student’s progress. MyST records the Phoenix parses for each student utterance, the dialogue rules triggered by the utterance, and the final state of the dialogue frames. These data used in conjunction with learning gains data are used to compute the Phoenix-based features and correlations in Table 8.2. The descriptions for these features can be found in Appendix C.2.

Of the Phoenix-based features PERCENT_PROMPTS_FROM_RULES had the largest correlation. Examination of the task files reveals artifacts of MyST dialogue authoring that may factor into this relationship. A majority of the rules in MyST task files are triggered by the filling of a slot and are used to deliver a QtA-style *Mark* or *Revoice* act. For example in one task frame, saying the word “circuit” will fill the [Circuit] slot, which will subsequently cause MyST to ask “Neat. I think I heard you say something about circuits. What are batteries all about in a circuit?”

These correlations may reflect two types of learning scenarios: 1) the student is triggering more rules because he understands the concepts and/or 2) the added MyST responsiveness by way of rules is causing a positive effect on learning. To further investigate this relation, students were binned into quartiles based on ASK pre-test score and were split into two groups. Re-plotting the percent of prompts from rules against the learning gains (Figure 8.3) shows a difference in the two populations. While the students in the upper two quartiles showed a positive correlation (R=0.13) between rules and learning, the students in the lower two quartiles showed an even stronger
correlation (R=0.36). This suggests that the higher correlation is not simply a matter of knowing the material, and the gains associated with the triggering of rules may have to do with the added interaction.

Across all students in the sample, the number of utterances that were unparseable by Phoenix (PERCENT_STUDENT_TURNS_UNPARSEABLE) is negatively correlated with learning. Conditioning this correlation on pre-test scores, shows different behavior between low and high-performing students (Figure 8.4). Students in the lower two pre-test score quartiles exhibit a negative correlation (R=-0.14) with residual gain, while students in the upper two quartiles exhibit a positive correlation (R=0.08) with learning. Phoenix’s inability to parse an utterance does not necessarily reflect the incorrectness or non-suitability of a student response. Rather, several conditions can lead to unparseable utterances including: unrecognized vocabulary, bad ASR output, or a slightly different syntactic construction. Furthermore, the Phoenix dialogue manager makes decisions only on its ability to parse. When given an unparseable utterance, the dialogue manager will default to asking the next prompt in the list. Thus the negative correlation among students in the lower two quartiles may actually reflect a correlation with lack of understanding. Conversely, the slightly positive correlation among students in the upper two quartiles may stem from using more advanced language or from having extra opportunities to express ideas.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>R</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERCENT_PROMPTS_FROM_RULES</td>
<td>0.20</td>
<td>0.09</td>
<td>0.23</td>
<td>0.01</td>
</tr>
<tr>
<td>PERCENT_SLOTS_FAILED</td>
<td>0.24</td>
<td>0.12</td>
<td>0.09</td>
<td>0.27</td>
</tr>
<tr>
<td>AVERAGE_TIMES_PROMPTED_PER_SLOT</td>
<td>1.00</td>
<td>0.39</td>
<td>0.01</td>
<td>0.89</td>
</tr>
<tr>
<td>PERCENT_STUDENT_TURNS_PARSE INTO_CURRENT_FRAME</td>
<td>0.35</td>
<td>0.14</td>
<td>0.01</td>
<td>0.89</td>
</tr>
<tr>
<td>PERCENT_STUDENT_TURNS_UNPARSEABLE</td>
<td>0.41</td>
<td>0.17</td>
<td>-0.08</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table 8.2: Phoenix Dialogue Manager Features Correlations: R is Pearson’s correlation coefficient between the feature and the residual learning gain, and p is the p-value for testing non-correlation. Rows are ordered by descending R values.
Frame: BatteryFunction
Description: The d cell is the source of electricity
[._start]+
  Action: "clear_screen(); synth(‘So, tell me more about batteries, what are they all about?’); send()"
[Battery]
[Source]+
  Action: "synth(‘What is important about the d cell when lighting a bulb?’); send()"
  Action: "clear_screen(); synth(‘Okay, tell me more what a battery or d cell, does in a circuit.’); send()"
  Action: "flash(‘(I)components’); synth(‘Roll over the d cell in this picture. You see the word d cell but it also says source. What is up with that?’); send()"
[Electricity]+
  Action: "clear_screen(); synth(‘Tell me more about what a d cell provides in a circuit.’); send()"
  Action: "flash(‘(I)wire chal 2’); synth(‘Here is a one wire circuit. What do you observe?’); send()"
  Action: "flash(‘(I)wire chal 2’); synth(‘Click on the plus sign to disconnect the d cell. How does this relate to what you think a d cell does in a circuit?’); send()"
[_done]+
  Action: "synth(‘So, in a circuit, the d cell is the source of electricity or energy for the circuit.’)"
Rules:
[Circuit] AND ![Source] AND ![Electricity]
  Action: "synth(‘Heat. I think I heard you say something about circuits. What are batteries all about in a circuit?’); send()"
[Source] AND [Source] != "FAIL" AND ![Electricity]
  Action: "synth(‘Interesting. You said something about the d cell being the source. What do you mean by that?’); send()"
  Action: "synth(‘You are correct in thinking that the d cell is the source of energy or electricity in a circuit. Good job!’); set(BatteryFunction:[_done] = ‘TRUE’)"

Frame: WiresFunction
Description: The wires carry electricity and can connect components
[._start]+
  Action: "clear_screen(); synth(‘Let’s talk about wires. What’s up with having wires in a circuit?’); send()"
[Wires]
[Carry]+
  Action: "synth(‘Tell me more about what the wires do in a circuit.’); send()"
  Action: "flash(‘(I)components’); synth(‘Roll over the green wires. You see the word pathway. What does it mean that wires are a pathway?’); send()"
[Electricity]+
  Action: "flash(‘(I)components’); synth(‘Tell me about what wires carry in a circuit.’); send()"
[_done]+
  Action: "flash(‘(I)components’); synth(‘So in a circuit, the wires carry electricity and can connect the components.’)"
Rules:

...
Figure 8.3: PERCENT_PROMPTS_FROM_RULES Versus Student Learning Gains, Conditioned on Pre-test Score.

Figure 8.4: PERCENT_STUDENT_TURNS_UNPARSEABLE Versus Student Learning Gains, Conditioned on Pre-test Score.

8.5.3 DISCUSS Features

Features based on DISCUSS allow a more in-depth analysis of the role of different tutoring moves and student responses in the learning process. As detailed above tutor turns were tagged with DISCUSS using a lookup table, while student turns were tagged using the DISCUSS label classifiers trained on the WOZ corpus. Descriptions of all the DISCUSS-based features are listed in Appendix C.3.
**Sequence Edit Distance Features:** Sequences of dialogue acts can reveal common patterns of behavior and can even be used to identify latent dialogue modes that correlate with learning (Boyer et al., 2009a). In practice this kind of discovery requires at least an order of magnitude more data and a more condensed dialogue move taxonomy. Though this is not feasible using DISCUSS, it is still possible to glean insight into the correlation between DISCUSS sequences and learning by calculating distances between extracted sequences.

A sequence of DISCUSS moves was extracted for each dialogue frame in a tutoring session. Because student moves are more variable and because the student utterance classifiers do not output a tuple structure, the sequences consist only of the DISCUSS tuples produced by MyST. Next the minimal sequence of DISCUSS tuples was extracted from the speak.discuss file. The minimal sequence is the smallest sequence of prompts needed to completely fill all the slots in the dialogue frame. To illustrate how these two sequences are converted to a feature for correlation consider the following example.

For the BatteryFunction frame listed in Figure 8.2 the minimal sequence of DISCUSS tuples $seq_{\text{min}}$ is:

1. **Ask/Elaborate/Entity** *(So, tell me more about batteries, what are they all about?)*
2. **Ask/Describe/Function** *(What is important about the d cell when lighting a bulb?)*
3. **Ask/Describe/Function** *(Tell me more about what a d cell provides in a circuit.)*
4. **Assert/Recap/Proposition** *(So, in a circuit, the d cell is the source of electricity or energy for the circuit.)*

and if the sequence extracted for a frame $seq_{\text{extract}}$ within a transcript is:

1. **Ask/Elaborate/Entity** *(So, tell me more about batteries, what are they all about?)*
2. **Ask/Describe/Function** *(What is important about the d cell when lighting a bulb?)*
3. **Ask/Describe/Function** *(Okay, tell me more what a battery or d cell does in a circuit)*
4. **Feedback/Positive/None** *(Neat.)*
5. **Mark/None/None** *(I think I heard you say something about circuits.)*
(6) Ask/Describe/Function (What are batteries all about in a circuit?)

(7) Assert/Recap/Proposition (So, in a circuit, the d cell is the source of electricity or energy for the circuit.)

Then a modified Levenshtein distance (Levenshtein, 1965) can be used to calculate the edit distance between \( seq_{min} \) and \( seq_{extract} \). The basic operations for computing a DISCUSS edit distance are insertions and deletions of Dialogue Acts, Rhetorical Forms, and Predicate Types with separate weights for each dimension. Qualitative analysis of clustering via edit distance found suitable costs for insertion and deletion to be 10, while the weights for the Dialogue Act, Rhetorical Form, and Predicate Type dimensions were 8, 4, and 8 respectively.

AVG\_DISTANCE\_TO\_MINIMAL\_FRAME\_SEQUENCE is then the average of all of the edit distances for all frames in the set of dialogues. Statistics for this feature are listed in Table 8.3. This feature was among the strongest and most negatively correlated of all features. A naïve interpretation of this result would assume students are best served with only a minimal sequence. Instead, the R of -0.26 reflects the difference in dialogues between struggling students and high performing students. Struggling students require additional prompts and scaffolding, while those who understand the material quickly fill the slots and complete the lesson.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>R</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVG_DISTANCE_TO_MINIMAL_FRAME_SEQUENCE</td>
<td>134</td>
<td>25.2</td>
<td>-0.26</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 8.3: Sequence Edit Distance Feature Correlations: R is Pearson’s correlation coefficient between the feature and the residual learning gain, and p is the p-value for testing non-correlation.

**Dialogue Act Features:** The Dialogue Act features were computed by counting the number of turns containing a given DISCUSS Dialogue Act (DA) tag, and dividing the total number of turns in a dialogue. Instead of calculating these percentages across all turns, they were conditioned on the speaker type (tutor versus student). Correlations between Dialogue Act features and residual learning gain percent are shown in Table 8.4.

For student turns, the dialogue acts with the largest magnitude in correlation appear to
contradict common intuitions. *MergedNoUnderstanding* and *Open* had a positive correlation with learning while *Answer* had a negative correlation. For *MergedNoUnderstanding* removing five outlier instances eliminated students that had both high learning gain scores and a tendency to respond “I don’t know” to every question. This resulted in a decreased Pearson correlation coefficient of $R=0.04$. The correlations for *Open* and *Answer* reflect two halves of the same phenomenon. It is not the case that the more a student answers the less he learns; rather the correlations are additional evidence that the better performing students are quicker to provide the answers to MyST. In that vein, the number of turns dedicated to greetings is close to a fixed amount. If the number of answers to complete a session decreases, then the percentage of turns with an *Open* dialogue act increases.

For tutor turns, the spectrum of features and correlations lend themselves to two possible interpretations. The first follows from the hypothesis that tutoring styles that encourage student self-expression and self-explanation lead to better learning than more direct or explicit approaches. The two dialogue acts with the largest (most positive) correlations with learning were *Feedback* and *Mark*. These two dialogue acts are hallmarks of the Questioning the Author style of teaching. Conversely the two smallest (most negative) correlations were with the *Assert* and *Direct* dialogue acts, suggesting again that excessive guidance from the tutor can stifle learning.

Another interpretation assumes these correlations reflect artifacts of MyST dialogue authoring. Differences in student abilities of expression lead to alternative dialogue states, which in turn expose issues in the authoring. For example the *Feedback* and *Mark* dialogue acts are prototypically used to highlight when a student states a key term: “Interesting! You said electricity. Tell me more about that.” QtA advocates using these moves regardless of student correctness; however a student who is unable to fill any of the dialogue slots is unlikely to receive tutor responses with these dialogue acts. Similarly, *Assert* and *Direct* are typically seen in MyST “bottom-out” conditions after repeated, unsuccessful attempts to elicit language from the student. An *Assert* prompt from the tutor may outright explain the concept MyST was attempting to elicit, and *Direct* are often used to guide the student in exploring the on-screen visuals. A similar explanation applies for the *Hint* dialogue act.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>R</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>STUDENT_TURNS_DA_MergedNoUnderstanding%</td>
<td>0.07</td>
<td>0.07</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>STUDENT_TURNS_DA_Open%</td>
<td>0.01</td>
<td>0.02</td>
<td>0.11</td>
<td>0.19</td>
</tr>
<tr>
<td>STUDENT_TURNS_DA_Acknowledge%</td>
<td>0.01</td>
<td>0.02</td>
<td>0.07</td>
<td>0.40</td>
</tr>
<tr>
<td>STUDENT_TURNS_DA_Metastatement%</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
<td>0.46</td>
</tr>
<tr>
<td>STUDENT_TURNS_DA_Thank%</td>
<td>0.00</td>
<td>0.01</td>
<td>−0.01</td>
<td>0.92</td>
</tr>
<tr>
<td>STUDENT_TURNS_DA_Close%</td>
<td>0.01</td>
<td>0.01</td>
<td>−0.02</td>
<td>0.83</td>
</tr>
<tr>
<td>STUDENT_TURNS_DA_Answer%</td>
<td>0.84</td>
<td>0.14</td>
<td>−0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Feedback%</td>
<td>0.16</td>
<td>0.05</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Mark%</td>
<td>0.03</td>
<td>0.02</td>
<td>0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Metastatement%</td>
<td>0.00</td>
<td>0.01</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Revoice%</td>
<td>0.03</td>
<td>0.02</td>
<td>0.05</td>
<td>0.52</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Acknowledge%</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.71</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Defer%</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.75</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Focus%</td>
<td>0.04</td>
<td>0.01</td>
<td>−0.02</td>
<td>0.79</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Hint%</td>
<td>0.03</td>
<td>0.02</td>
<td>−0.10</td>
<td>0.25</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Ask%</td>
<td>0.70</td>
<td>0.06</td>
<td>−0.10</td>
<td>0.22</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Assert%</td>
<td>0.20</td>
<td>0.03</td>
<td>−0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>TUTOR_TURNS_DA_Direct%</td>
<td>0.15</td>
<td>0.05</td>
<td>−0.31</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 8.4: Dialogue Act Feature Correlations: R is Pearson’s correlation coefficient between the feature and the residual learning gain, and p is the p-value for testing non-correlation. Features have been broken down by student turns and tutor turns. Rows are sorted by descending R-values.

**Rhetorical Form Features:** Like with dialogue act features, rhetorical form features are computed as the proportion of speaker turns containing a tag. Correlation statistics between rhetorical forms and residual learning gain are listed in Table 8.5. From a pedagogical perspective, the high correlation between *Justify* and learning gains aligns with the Questioning the Author philosophy of encouraging the student to express concepts in his or her own language. Eliminating instances where the *Justify* percentage equals zero, essentially eliminates non-relevant lessons and boosts the Pearson correlation coefficient to R=0.45. For the positively correlated *List* and the negatively correlated *Quantify* there is no immediate explanation, though one could argue that the process of listing leads to further explanation whereas a student may not think beyond the literal answer requested making a quantification. Explanations for other correlations follow directly from the dialogue act explanations above when considered in context of common dialogue act/rhetorical form pairings such as *Direct/Attend*, *Direct/Task*, and *Feedback/Positive*. 
<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>R</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>STUDENT_TURNS_RF_List%</td>
<td>0.12</td>
<td>0.06</td>
<td>0.25</td>
<td>0.00</td>
</tr>
<tr>
<td>STUDENT_TURNS_RF_Justify%</td>
<td>0.05</td>
<td>0.05</td>
<td>0.18</td>
<td>0.03</td>
</tr>
<tr>
<td>STUDENT_TURNS_RF_Greet%</td>
<td>0.01</td>
<td>0.02</td>
<td>0.10</td>
<td>0.25</td>
</tr>
<tr>
<td>STUDENT_TURNS_RF_Bye%</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.76</td>
</tr>
<tr>
<td>STUDENT_TURNS_RF_MergedYesNo%</td>
<td>0.61</td>
<td>0.18</td>
<td>-0.03</td>
<td>0.70</td>
</tr>
<tr>
<td>STUDENT_TURNS_RF_Compare%</td>
<td>0.05</td>
<td>0.05</td>
<td>-0.04</td>
<td>0.68</td>
</tr>
<tr>
<td>STUDENT_TURNS_RF_Identify%</td>
<td>0.14</td>
<td>0.07</td>
<td>-0.07</td>
<td>0.38</td>
</tr>
<tr>
<td>STUDENT_TURNS_RF_Quantity%</td>
<td>0.03</td>
<td>0.04</td>
<td>-0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_Justify%</td>
<td>0.01</td>
<td>0.02</td>
<td>0.28</td>
<td>0.00</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_Positive%</td>
<td>0.16</td>
<td>0.05</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_List%</td>
<td>0.02</td>
<td>0.02</td>
<td>0.24</td>
<td>0.00</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_Recap%</td>
<td>0.11</td>
<td>0.02</td>
<td>0.11</td>
<td>0.18</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_Quantity%</td>
<td>0.01</td>
<td>0.01</td>
<td>0.08</td>
<td>0.34</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_Select%</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.97</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_Meta%</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.05</td>
<td>0.59</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_Compare%</td>
<td>0.04</td>
<td>0.02</td>
<td>-0.07</td>
<td>0.40</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_Identify%</td>
<td>0.03</td>
<td>0.03</td>
<td>-0.07</td>
<td>0.38</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_Task%</td>
<td>0.09</td>
<td>0.03</td>
<td>-0.22</td>
<td>0.01</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_MergedDescribe%</td>
<td>0.64</td>
<td>0.07</td>
<td>-0.25</td>
<td>0.00</td>
</tr>
<tr>
<td>TUTOR_TURNS_RF_Attend%</td>
<td>0.06</td>
<td>0.02</td>
<td>-0.32</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 8.5: Rhetorical Form Feature Correlations: R is Pearson’s correlation coefficient between the feature and the residual learning gain, and p is the p-value for testing non-correlation. Features have been broken down by student turns and tutor turns. Rows are sorted by descending R-values.

**Predicate Type Features:** Predicate type features are computed in the same way as the dialogue act and rhetorical form features. Table 8.6 contains the correlation statistics between predicate type percentages and residual learning gains.

With predicate types the two common themes of authoring artifacts and useful dialogue properties re-emerge. The negatively correlated *MergedVisual* and *MergedCausalRelation* features are among the most frequently used tags. Their inverse relation with learning largely added additional questioning from MyST which is caused when dialogue slots are not filled. On the other hand, the more positively correlated features such as *Procedure* and *Entity* and *MergedTopic* may be more indicative of the utility of asking more open-ended questions that force the student to describe either what they did in class or what they see on screen. Conditioning these correlations
<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>R</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>STUDENT_TURNS_PT_Procedure%</td>
<td>0.18</td>
<td>0.11</td>
<td>0.25</td>
<td>0.00</td>
</tr>
<tr>
<td>STUDENT_TURNS_PT_MergedTopic%</td>
<td>0.13</td>
<td>0.06</td>
<td>0.19</td>
<td>0.02</td>
</tr>
<tr>
<td>STUDENT_TURNS_PT_Function%</td>
<td>0.20</td>
<td>0.12</td>
<td>0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>STUDENT_TURNS_PT_MergedYesNoMaybe%</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.96</td>
</tr>
<tr>
<td>STUDENT_TURNS_PT_Entity%</td>
<td>0.21</td>
<td>0.08</td>
<td>−0.03</td>
<td>0.75</td>
</tr>
<tr>
<td>STUDENT_TURNS_PT_MergedCausalRelation%</td>
<td>0.26</td>
<td>0.19</td>
<td>−0.03</td>
<td>0.70</td>
</tr>
<tr>
<td>STUDENT_TURNS_PT_MergedRoute%</td>
<td>0.05</td>
<td>0.09</td>
<td>−0.06</td>
<td>0.48</td>
</tr>
<tr>
<td>STUDENT_TURNS_PT_MergedVisual%</td>
<td>0.25</td>
<td>0.12</td>
<td>−0.18</td>
<td>0.03</td>
</tr>
<tr>
<td>TUTOR_TURNS_PT_Entity%</td>
<td>0.07</td>
<td>0.03</td>
<td>0.19</td>
<td>0.02</td>
</tr>
<tr>
<td>TUTOR_TURNS_PT_Procedure%</td>
<td>0.11</td>
<td>0.07</td>
<td>0.16</td>
<td>0.05</td>
</tr>
<tr>
<td>TUTOR_TURNS_PT_MergedTopic%</td>
<td>0.13</td>
<td>0.03</td>
<td>0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>TUTOR_TURNS_PT_Proposition%</td>
<td>0.11</td>
<td>0.02</td>
<td>0.05</td>
<td>0.57</td>
</tr>
<tr>
<td>TUTOR_TURNS_PT_Attribute%</td>
<td>0.15</td>
<td>0.08</td>
<td>0.03</td>
<td>0.71</td>
</tr>
<tr>
<td>TUTOR_TURNS_PT_Function%</td>
<td>0.07</td>
<td>0.04</td>
<td>−0.01</td>
<td>0.94</td>
</tr>
<tr>
<td>TUTOR_TURNS_PT_Configuration%</td>
<td>0.02</td>
<td>0.03</td>
<td>−0.07</td>
<td>0.41</td>
</tr>
<tr>
<td>TUTOR_TURNS_PT_MergedRoute%</td>
<td>0.02</td>
<td>0.02</td>
<td>−0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>TUTOR_TURNS_PT_MergedCausalRelation%</td>
<td>0.11</td>
<td>0.09</td>
<td>−0.19</td>
<td>0.02</td>
</tr>
<tr>
<td>TUTOR_TURNS_PT_MergedVisual%</td>
<td>0.22</td>
<td>0.07</td>
<td>−0.24</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 8.6: Predicate Type Feature Correlations: R is Pearson’s correlation coefficient between the feature and the residual learning gain, and p is the p-value for testing non-correlation. Features have been broken down by student turns and tutor turns. Rows are sorted by descending R-values.

on pre-test score reveals interesting contrasts between the students who presumably understand the material a priori and those who do not. The learning gain correlation for students in the lower two pre-test quartiles are $R_{Procedure} = 0.31$ and $R_{MergedTopic} = 0.29$, and for students in the upper two quartiles $R_{Procedure} = 0.08$ and $R_{MergedTopic} = −0.03$. This suggests that the act of verbalizing and rehashing classroom labs and activities is especially important for students who do not fully grasp the material.

**DISCUSS Tuple Features:** The full DISCUSS tuple from the tutors’ turns give additional detail about the specific form of the tutor’s prompts. Because of both the combinatoric induced sparsity and the limited number of samples, the tuple analysis does not provide any insight beyond what was already detailed in the separate DISCUSS dimensions. Looking at the largest tuple correlations (Table 8.7) reconfirms the results from above.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>R</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUTOR_TURNS_DARFPT_Ask_Justify_Procedure%</td>
<td>0.01</td>
<td>0.02</td>
<td>0.29</td>
<td>0.00</td>
</tr>
<tr>
<td>TUTOR_TURNS_DARFPT_Ask_List_Entity%</td>
<td>0.01</td>
<td>0.01</td>
<td>0.29</td>
<td>0.00</td>
</tr>
<tr>
<td>TUTOR_TURNS_DARFPT_Feedback_Positive_None%</td>
<td>0.16</td>
<td>0.05</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td>TUTOR_TURNS_DARFPT_Ask_Recap_MergedTopic%</td>
<td>0.00</td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>TUTOR_TURNS_DARFPT_Ask_MergedDescribe_Proposition%</td>
<td>0.00</td>
<td>0.00</td>
<td>0.22</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 8.7: DISCUSS Tuple Feature Correlations: R is Pearson’s correlation coefficient between the feature and the residual learning gain, and p is the p-value for testing non-correlation. In the interest of space, only the five most correlated and five most negatively correlated tuple features are listed.

8.6 Conclusions and Future Directions

This chapter has illustrated how features extracted from the transcripts log files can be used to analyze the nature of dialogue interactions within an intelligent tutoring system. The results for shallow dialogue features reconfirm many of the findings in the literature regarding student and tutor behavior. Adding Phoenix and DISCUSS features provides a more detailed representation for examining the form and makeup of tutor and student responses, and provides insight into the aspects of the QtA pedagogy that are most beneficial to learning within the MyST environment. More importantly the use of rich semantic and pragmatic representations such as DISCUSS can serve as a general, lesson-independent framework for more directed inquiries into the nature of tutoring. Proper experimental design and application of this framework can provide an empirical methodology and statistical justification for design, modifications and refinements to dialogue within intelligent tutoring systems.

With respect to the MyST assessment data, the results provide a first order verification of the pedagogical principles and practices that informed MyST dialogue design. While this does not establish strong causal ties, it does demonstrate how DISCUSS features add information that could not be gleaned from shallow features alone. Moreover the correlation statistics highlight artifacts
and biases about the dialogue that the system designers and dialogue authors might otherwise overlook. For example, core Questioning the Authoring tactics such as feedback and marking are primarily used in conditions that benefit students who are already performing at or above average, while struggling students may never encounter these moves. This is a byproduct of how most task files are authored, and it would be interesting to see if the positive correlations between these acts and learning gains still hold if applied in other contexts.

The largest limitations in this study were the lack of variability in MyST dialogue design and the coarseness of the ASK assessment data. While these decisions were sufficient for the purpose of demonstrating the effectiveness of a conversational virtual tutor for elementary school science education, they were not well suited for teasing apart differences in questioning tactics, student behavior and tutoring strategies. Establishing stronger, more conclusive correlations will require an experimental design that introduces more variation in the type of questions and strategies employed by the tutoring system. Tighter coupling between assessment and dialogue would also help to ensure the sample is more uniform in its distribution of features.

One possible follow-up experiment would be to rewrite the MyST task files in a way to keep the underlying DISCUSS representation intact while simultaneously shifting to a more direct tutoring style than QtA. Running an assessment study on students from direct-tutoring and QtA-tutoring conditions would help to identify the relative merits and drawbacks of each approach. With enough data, correlation analyses could help to determine what dialogue behaviors are best suited for each style. To explore the range and benefit of tactics within a single pedagogy would require authoring more alternative dialogue paths. In the short term, DISCUSS dimensions could define the space of possible variants. Looking forward, improvements in question generation technology could provide the foundation for more systematic control and manipulation of the dialogue.
Chapter 9

Conclusions and Future Work

The work presented in this thesis is a step toward the larger goals of enabling data-driven creation and refinement of task-oriented dialogue systems and of establishing an experimental framework that supports inquiry into the strengths and weakness of various tutoring and pedagogical strategies. Yet, substantial work remains along the path to these greater goals. Sections 9.1 and 9.2 help to situate this work’s contributions with respect to these greater aims, while Section 9.3 discusses the lessons learned and their ramifications for future work. Lastly, this thesis concludes with some closing remarks in Section 9.4.

9.1 Summary

The major contributions of this work are 1) defining the Dialogue Schema Unifying Speech and Semantics (DISCUSS), a semantically and pragmatically informed representation of dialogue that abstracts speakers’ moves to their underlying action, function and content, 2) creating a corpus of tutorial dialogues annotated with this DISCUSS representation, 3) developing statistical and machine learning-based models to automatically annotate utterances with DISCUSS tags, and 4) demonstrating the utility of DISCUSS in two tutorial dialogue-oriented tasks: question selection and learning gains prediction. These contributions show the feasibility of annotating complex dialogue representations such as DISCUSS on real human-human and human-computer interactions. Results from the various dialogue tasks presented in this thesis illustrate the additional insights and performance that can be gained for dialogue automation and analysis. Together these contributions
provide a framework for additional inquiry into the nature of human-computer dialogue as well as tutor and student behavior within intelligent tutoring systems.

### 9.2 Research Questions Revisited

**Research Question 1.** What semantic and pragmatic representations are necessary for modeling tutorial dialogue and at what granularity?

**Hypothesis 1.1** Dialogue acts alone are too coarse to capture the types of prompts and responses seen in tutoring. Conversely, sentence-level semantic representations are too specific for modeling conversational discourse. Additional syntactico-semantic information is needed to bridge this gap and to imbue dialogue moves with a more complete account of the action, function, and content contained within a dialogue.

**Hypothesis 1.2** Learning domain-independent dialogue behaviors from data will require a way to generalize the task-specific semantics.

- A survey of existing dialogue act, tutoring act, and question-type taxonomies led to the creation of the Dialogue Schema for Speech and Semantics (DISCUSS) a dialogue move taxonomy that layers dialogue and rhetorical annotation over semantic representations. DISCUSS’s dialogue act, rhetorical form, and predicate type dimensions provides a factorization that accounts for the semantics and pragmatics of dialogue. Analysis of results in dialogue-automation and dialogue-analysis tasks shows that features derived from each of DISCUSS’s dimensions contribute to system performance.

**Research Question 2.** Are DISCUSS categories well enough defined to attain usable inter-annotator agreement?

**Hypothesis 2.1** Based on existing literature on dialogue move annotation, linguists (with sufficient training) should show high inter-annotator agreement on dialogue act annotation.
While annotating discourse moves and semantic predicate categories offer more opportunities for confusion, annotators should achieve fair but usable inter-annotator agreement. (Chapter 5)

- Inter-annotator agreement computed from a corpus of DISCUSS-tagged tutorial dialogues found kappa statistics of 0.76, 0.72, and 0.63 for the dialogue act, rhetorical form, and predicate type dimensions respectively. These levels give fair confidence in the labeling task reliability and suggest that training models to replicate human-annotation is feasible.

**Hypothesis 2.2** Computational models trained to label dialogues with a rich dialogue move taxonomy should achieve results in-line with inter-annotator agreement scores.

- Automatic DISCUSS classifiers achieved frequency-weighted $F_1$-scores of 0.935 for dialogue acts, 0.704 for rhetorical forms, and 0.530 for predicate types, using commonly used NLP features, which follow directly from the inter-annotator reliability statistics. Commonly confused classes were merged to improve decision boundaries and to create more meaningful, reliable DISCUSS categories. Classifiers trained to label using the merged classes showed increased performance over the original unmerged classes. These results demonstrate the feasibility of automatic DISCUSS labeling for student utterances.

**Research Question 3** How do DISCUSS-based features aid in dialogue decision making tasks such as selecting follow-up questions?

**Hypothesis 3.1** Questioning styles and preferences vary from tutor to tutor, even when teaching with the same pedagogical philosophy. Modeling these individual preferences requires a range of features from shallow lexical features to more complex semantic, pragmatic, and dialogue context features. (Chapter 7)

- A system trained to rank and select follow-up questions within the course of a dialogue exhibited performance and agreement on par with experienced human tutors. Adding DISCUSS-derived features illustrates the added predictive power gleaned from a syntactico-semantic dialogue move representation. Inspection of feature weights and distributions shows that the diversity of features afforded by the dialogue act, rhetorical
form, and predicate type layers allows for better modeling of individual preferences in question asking.

**Research Question 4** What additional insights can a rich, dialogue move representation like DISCUSS provide for dialogue analysis tasks such as predicting learning gains from dialogue transcripts?

**Hypothesis 4.1** While shallow measures such as student utterance length have been shown to correlate with learning, exploring more detailed phenomena like the role of self-explanation in learning requires more detailed information about the kinds of questions tutors ask and the answers students provide. (Chapter 8)

- Correlating DISCUSS-derived features with measured learning gains provided more detailed insights into the types of behaviors that align with positive and negative learning experiences than could be attained with shallow features alone. These correlations confirmed the usefulness of the pedagogical principles and practices driving MyST system design and helped to highlight artifacts of systems dialogue authoring that need addressing. Most importantly, this representation was able to capture the strengths and weaknesses of the QtA-style employed in MyST.

**9.3 Discussion and Future Work**

Though this thesis has shown the benefits of dialogue taxonomies such as DISCUSS for automating and analyzing dialogues, it is neither perfect nor final. The addition of syntactico-semantic annotation layers like the rhetorical form and predicate type dimensions greatly adds to the descriptive power of DISCUSS over more coarse dialogue taxonomies like DAMSL; however there are still numerous linguistic phenomena to capture to make for truly robust human-computer dialogue experiences. At this time DISCUSS largely ignores task and dialogue structure. While initial versions of DISCUSS included the notion of links to indicate how turns related to each other and to the goals of the dialogue, this proved to be too difficult of a task for the annotators. Inclusion
of discourse and dialogue structure could be used for incremental parsing of dialogue (Bangalore and Stent, 2009), which in turn can add richer context features for making dialogue decisions.

There is still much work to improve DISCUSS inter-annotator agreement from merely usable to highly reliable. Because of sparsity, it may be more prudent to modify the annotation task to more closely mirror the binary tagging task. Instead of enforcing strict DISCUSS tuples, it may be more practical for annotators to decide whether or not an utterance exhibits any of several DISCUSS properties. Separating lexical cues from conversational intent is another recurring challenge for DISCUSS annotation, especially for the ambiguous predicate type dimension. Annotators would often conflate lexical items for semantic meaning. For example the utterances containing the word “pathway” were often tagged with predicate type Route regardless of whether or not a trajectory or direction was expressed. While some of this can be overcome with additional annotator training and more specific guidelines, predicate type annotation may be more successful if it was grounded in relation to other lexical resources such as VerbNet (Schuler, 2005) or FrameNet (Baker et al., 1998). While FrameNet and VerbNet look closely at the sentence level semantics, a predicate type can be regarded as a way of organizing frame categories. Additionally, it would be worthwhile to evaluate the use of DISCUSS on different genres or even on transcripts featuring a different tutoring style. Questioning the Author derives much of its educational power from using deliberately vague prompts to encourage student understanding. This vagueness was the source of several annotation problems as the annotators were unable to discern the intent or even the subject of the tutor’s prompts.

Another challenge lies in scaling out DISCUSS classifiers for use across multiple lessons, subjects, and domains. Currently, the automatic DISCUSS classifiers rely heavily on lexical features, which inherently come back to lesson-specific words – especially for the more semantically-oriented predicate type tags. While additional annotation effort could provide enough data to reduce reliance on lexical features and shift toward more sophisticated syntactic features, this alone may not account for the signatures necessary for detecting syntacto-semantic categories like DISCUSS predicate types. Alternatively, classification of DISCUSS tags may better generalize via features based on semantic
information extracted from the dialogue system’s natural language understanding unit. However, for current systems, including MyST/Phoenix, these frames are still too domain-specific to allow for this generalization. One possible first step would be to map elements in the existing frames to VerbNet categories. Looking out further, authoring dialogues around more domain-independent frames may prove to be a necessary step to achieving robust generalization of dialogue tactics.

The results from the experiments in ranking and selecting follow-up questions are promising and offer a path toward data-driven automation of dialogue behavior. In addition to illustrating the importance of DISCUSS features for performing the task, the methodology and approaches used to gather and utilize data highlight DISCUSS’s potential as a representation for natural language generation. While the experiments in this work made use of manually authored questions, a question generator that systematically generates questions via permutation of DISCUSS tags would allow for more controlled experimentation. Managing the variation in question categories would provide an even more principled way of identifying and learning the factors that drive decision making during tutoring.

The inquiries into correlations with learning gains demonstrated the usefulness of DISCUSS for describing and detecting the interaction signatures found within MyST tutorial dialogues. As noted in Chapter 8, the largest challenges to discovering strong correlations were caused by the 1) loose coupling between the learning gain score and the individual dialogue sessions and 2) the lack of variation in dialogue authoring. Because the MyST learning gains study was more focused on the MyST approach as a whole, the collected data were not at the ideal granularity for correlating dialogue features with learning. Future experiments that seek to better understand the connection between dialogue behaviors and learning would benefit from a more focused experimental design. Within a FOSS framework, such inquiries may be better answered by limiting the set of dialogues to a single investigation. Additionally, tying assessment tests to individual tutoring sessions would give more direct insight into these phenomena. At the same time, special care should be taken to vary dialogue tactics. With the current authoring, students who get correct answers fall into one path while students who struggle follow a different one. Consequently each category of student
only experiences a subset of the question types. Like with question generation, more systematic variation of tutoring tactics could help to establish better correlations. Furthermore, with a refined experimental design and enough data, DISCUSS-based models could develop into new tools for assisting tutors and educators in assessing and evaluating their own teaching.

9.4 Closing Remarks

This dissertation was guided by the vision of one day developing and deploying tutoring systems capable of conversing with students and scaffolding them through any subject imaginable. Access to personalized tutoring at this scale could revolutionize education and greatly aid teachers who are continually faced with increasing class sizes and decreasing budgets. Improvements in tutorial dialogue technologies present a path toward providing every student with an engaging learning experience that enables individual ownership and discovery of knowledge. Progress toward these goals largely hinges on the ability to learn, generalize, and automate more robust and more intelligent dialogue behaviors. The relatively recent convergence of machine learning, natural language processing, and education has presented a new lens in which to explore the linguistic phenomena underlying tutorial dialogue interactions. With such tools at our collective disposal, the momentum and timing are perfect for making intelligent tutoring systems available for every learner. This dissertation is a step towards turning this vision of intelligent tutoring systems into reality.
Bibliography


Appendix A

DISCUSS Annotation Guidelines

A.1 Dialogue Act Definitions

A.1.1 Dialogue Control Tags

A.1.1.1 Acknowledge

This is a grounding act used by the speaker to signal understanding or reception of a previous utterance. Typical forms include repetition or back-channel response.

a) **Tutor**: very good, so the magnetism is going through the nail. we call this temporary magnetism

   (Revoice/–/–)

   **Student**: very good, so the magnetism is going through the nail. we call this temporary magnetism

   (Acknowledge/–/–)

A.1.1.2 Apologize

This tag marks an utterance containing an apology.

a) **Student**: sorry marnie it’s time for me to go bye

   (Apologize/–/–, Close/Bye/–)
A.1.1.3 Close

This dialogue acts marks moves that close a conversation. This is typically done with some version of a goodbye.

a) Tutor: Thats all the time we have for today
       (Close/–/–)

b) Tutor: Goodbye.
       (Close/Bye/–)

A.1.1.4 Metastatement

This marks utterances that reflect on the task, a personal state, or the state of the dialogue. It is not relay information regarding the topic under discussion.

a) Student: Hold on. I'm thinking
       (Metastatement/–/–)

A.1.1.5 Open

This dialogue acts denotes entries or introductions into a conversation or dialogue. Typically this is a greeting, but a conversation could open without a greeting as well.

a) Tutor: Welcome to the tutoring session
       (Open/–/–)

b) Tutor: Hi, how's it going?
       (Open/Greet/–)

A.1.1.6 Repeat, RequestRepeat

A Repeat is restatement of something stated previously, whereas a RequestRepeat is asking the other speaking to do a repeat.
a) Tutor: Tell me more about magnets

(Ask/Elaborate/Topic)

Student: Could you say that again?

(RequestRepeat/–/–)

Tutor: I said tell me more about magnets

(Repeat/–/–)

A.1.1.7 SignalNoUnderstanding

Signals a lack of understanding or knowledge

a) Tutor: What is the relation between the force and the distance?

(Ask/Describe/CausalRelation)

Student: I don’t know

(SignalNoUnderstanding/–/–)

A.1.2 Information Exchange Tags

A.1.2.1 Assert

A statement or an assertion made to give information to the other speaker(s). An Assert unlike an Answer provides an unsolicited statement. In practice, tutors often use this to provide the student with information relating to the lesson plan or to recap a key concept.

a) Tutor: The lightbulbs in a series circuit share a single pathway of electricity

(Assert/Describe/Configuration)

b) Tutor: Nice observations. Magnetism is a force that can go through space and other materials, like the paper that this box is made out of

(Feedback/Positive/–, Assert/Recap/Proposition)

A.1.2.2 Ask

Request information from another speaker, often in the form of a question.
a) **Tutor:** Tell me what’s going on in this picture

(Ask/Describe/Visual)

b) **Tutor:** Tell me which one is the battery

(Ask/Identify/Entity)

### A.1.2.3 Answer

The complementary act to Ask. Used to give information in response to a request.

a) **Tutor:** What did you do in science today?

(Ask/Describe/Activity)

b) **Student:** We measured a bunch of gravel and an apple and we learned about grams.

(Answer/Describe/Activity)

### A.1.2.4 Hint

In a Hint the tutor is providing a relevant fact or bit of information to help push the student towards expressing the correct answer.

a) **Tutor:** Look closely at the ends of the battery and then tell me again about the direction of flow of electricity

(Hint/–/–)

### A.1.2.5 Mark

A Mark is used to highlight or point out specific words or phrases from the other speakers past turn. Typically used by a tutor to draw attention to important concepts, and often used in conjunction with a follow up question.
a) **Tutor:** What do you notice about this magnet?
   
   (Ask/Describe/Observation)

   **Student:** It’s sticking to things
   
   (Answer/Describe/Observation)

   **Student:** You said sticking. Interesting. Can you tell me more about the sticking?
   
   (Mark/--/, Ask/Elaborate/Observation)

A.1.2.6 Revoice

Revoice is similar to a Mark, but instead of highlighting words or phrases from an utterance, the speaker summarizes or paraphrases it. This is often used by a tutor to help bring clarity to a concept to then allow the conversation to move forward.

a) **Tutor:** Think more about the magnets. What is up with magnets that make them attract or repel each other?
   
   (Ask/Describe/Process)

   **Student:** the attract by the sides the if they’re like if they’re a if they have a different side if you flip the magnet that attracts the other one and if you flip to the other way they’re not attracting because the other side has metal or another type of of thing not iron or steel but the other side has iron or and the other part doesn’t
   
   (Answer/Describe/Process)

   **Tutor:** I think you are talking about the fact that magnets have two sides or poles. Tell me more about that.
   
   (Revoice/--/, Ask/Elaborate/Configuration)

A.1.2.7 Recall

Indicates this utterance makes reference to something from the past (intra or extra-conversational).
a) **Tutor:** Earlier we observed magnets sticking to certain metals. Tell me more about that.
   (Recall/-/Experience, Ask/Elaborate/Experience)

A.1.3 Attention Management Tags

A.1.3.1 Defer

Indicates to the other speaker that the topic should be abandoned or saved for another occasion.

a) **Tutor:** I think you said something about the direction of flow. We will talk more about that in just a few minutes.
   (Mark/-/-, Defer/-/Topic)

A.1.3.2 Focus

A marked shift to switch topic or subject matter.

a) **Tutor:** Let’s think again about magnets. What is going on between them that makes them stick together?
   (Focus/-/Topic, Ask/Describe/Process)

A.1.3.3 Direct

Used to give instructions or directions to the other speaker.

a) **Tutor:** Try clicking on the battery.
   (Direct/Task/Visual)

b) **Tutor:** Look at this picture.
   (Direct/Attend/Visual)
A.2 Rhetorical Form Definitions

A.2.1 Dialogue Control Tags

A.2.1.1 Bye

Used in conjunction with the Close dialogue act to indicate if a goodbye occurred.

a) Tutor: That’s all for today. Talk to you next time.

   (Close/Bye/–)

A.2.1.2 Greet

Used in conjunction with the Open dialogue act to indicate if the opening has a hello-esque form.

a) Tutor: Hi, how’s it going?

   (Open/Greet/–)

A.2.1.3 Positive, Negative

Used in conjunction with the Feedback dialogue act to indicate the polarity of the feedback.

a) Tutor: Good thinking!

   (Feedback/Positive/None)

b) Tutor: That doesn’t sound right.

   (Feedback/Negative/None)

A.2.2 Information Exchange Tags

A.2.2.1 Clarify

Used either when asking the other speaker to clarify a position or when giving a clarification.
a) **Student:** well one picture shows that the light bulb has lots of energy and the other light bulbs don’t have that much energy as the one that has lots of energy

(Answer/Describe/Observation)

**Tutor:** What do you mean by something having less energy? What do you observe that makes you say that?

(Ask/Clarify/Observation, Ask/Justify/Observation)

### A.2.2.2 Compare

Used either when asking for or offering a comparison.

a) **Tutor:** Tell me about the nails. How are they alike or different?

(Ask/Compare/Attribute)

**Student:** they’re different because the aluminum nail doesn’t have any iron or steel in it and the steel nail has steel in it

(Answer/Compare/Attribute)

### A.2.2.3 Confirm

Confirms something stated in a previous utterance typically in the form of yes or no. This is more specific than a YesNo question, as it is looking to verify some already mentioned piece of knowledge.

a) **Tutor:** Did you talk about measurement in class today?

(Ask/Confirm/Activity)

**Student:** nope

(Answer/Confirm/AcceptRejectMaybe)

### A.2.2.4 Define

Used either when asking for or offering a definition.
a) **Tutor:** What does repel mean?

(Ask/Define/Topic)

**Student:** Repel means that the magnets push away from each other

(Answer/Define/Topic)

A.2.2.5 Describe

Used either when asking for or offering a description. Typically this is used to convey concepts or propositional content.

a) **Tutor:** What is up with them that makes magnets attract or repel?

(Ask/Describe/CausalRelation)

**Student:** If the same poles are facing each other they repel.

(Answer/Describe/Topic)

A.2.2.6 Elaborate

Used either when asking for or offering an elaboration. This is similar to the Describe rhetorical form, but with a more specific context that where previously establish information is expounded upon.

a) **Tutor:** Tell me more about the sides or ends of a magnet.

(Ask/Elaborate/Configuration)

**Student:** They have labels that are s and n which stand for south and north

(Answer/Describe/Topic)

A.2.2.7 Identify

Used either when asking for or offering an identification. This is often used to name or point out objects in pictures.
a) **Tutor:** *Which one is the battery?*  
(Ask/Identify/Entity)  

**Student:** *The thing with the plus and minus on it is a battery*  
(Answer/Identify/Entity)

b) **Tutor:** *The object on the left side is a magnet*  
(Assert/Identify/Entity)

### A.2.2.8 Justify

Used either when asking for or offering some form of justification. When used with an *Ask* dialogue act, the question is trying to elicit the reasons behind some evidence, otherwise it is to give support for something previously stated or observed.

a) **Student:** *That's a parallel circuit.*  
(Answer/Identify/Configuration)  

**Tutor:** *What makes you think that?*  
(Ask/Justify/Configuration)  

**Student:** *Because there are two wires.*  
(Answer/Justify/Configuration)

### A.2.2.9 List

Used to list entity mentions. This can be like *Identify* but for multiple items.

a) **Tutor:** *What were the things you used?*  
(Ask/List/Entity)  

**Student:** *We used a battery, wires, and a lightbulb.*  
(Answer/List/Entity)

### A.2.2.10 Predict

Used either when asking for or giving a prediction.
a) Tutor: What do you think will happen if we close the switch
   (Ask/Predict/CausalRelation)

Student: The electricity will stop flowing
   (Answer/Predict/CausalRelation)

A.2.2.11 Quantify

Used either when asking for or giving some form of quantification.

a) Tutor: How many pathways are there in a series circuit?
   (Ask/Quantify/Entity)

Student: There are two pathways in a series circuit.
   (Answer/Quantify/Entity)

A.2.2.12 Recap

Used to review or summarize what was previously said. An Assert/Recap/Proposition triplet differs slightly from a Revoice act as a recap is restating information that was held mutually by all parties, whereas a revoice is used to refine a student statement to drive the conversation forward or to bring clarity.

a) Tutor: I think you are really figuring this out. Magnets stick to objects that are made of iron or steel.
   (Assert/Recap/Proposition)

A.2.2.13 Select

Used for presenting or selecting from a choice.

a) Tutor: Will the magnet stick to the nail or the button?
   (Ask/Select/Entity)

Student: The nail.
   (Answer/Select/Entity)
A.2.2.14   YesNo

Unlike the Confirm rhetorical form, which is used to verify, YesNo is used to ask about or assert the truth or false condition of some proposition.

a)  Tutor:   Does it matter how far apart the magnets are?  
( Ask/YesNo/Configuration)

    Student:   Yeah.
    ( Answer/YesNo/YesNoMaybe)

A.2.3   Attention Management Tags

A.2.3.1   Attend

An Attend rhetorical form denotes that something should be addressed or dealt with. In contrast to focus, this is less about topic, and more about where attention should be placed (possibly physically). Typically accompanied by a Direct dialogue act.

a)  Tutor:   Look at these two circuits
( Direct/Attend/Visual)

A.2.3.2   Meta

The Meta rhetorical form indicates that the utterance references the system or provides some other self-referential commentary not directly related to the task or information under discussion. This is different from the Metastatement dialogue act because of its association with the Direct dialogue act. Whereas Metastatement is extra-conversational, the Meta is used to show that the instruction or direction is extra-topical.

a)  Tutor:   Think carefully.
( Direct/Meta/-)
A.2.3.3 Task

Used to indicate that the utterance is in reference to a task of learning activity. Typically used with the Direct dialogue act to give directions related to some activity.

a) Tutor: \textit{Try clicking on the magnets} \\
\quad (Direct/Task/Visual)

A.3 Predicate Type Definitions

A.3.0.4 AcceptRejectMaybe

Used for accepting, rejecting, or hedging in response to a request for confirmation.

a) Tutor: \textit{Did you say that you played with magnets?} \\
\quad (Ask/Confirm/Experience)

Student: \textit{Yeah we did} \\
\quad (Answer/Confirm/AcceptRejectMaybe)

A.3.0.5 Activity

Used to indicate that the utterance references some activity the speaker may have engaged in as well as a current activity under discussion.

a) Tutor: \textit{What have you been doing in science?} \\
\quad (Ask/Describe/Activity)

Student: \textit{We played with magnets} \\
\quad (Answer/Describe/Activity)

A.3.0.6 Attribute

Used when describes some property or attribute of the agent, instrument, theme or other semantic role in the sentence.
a) Tutor: Tell me more about the kinds of materials that stick to magnets.

(Ask/Elaborate/Attribute)

Student: Ones made of steel and iron

(Answer/List/Attribute)

A.3.0.7 CausalRelation

Used to describe relationships that have a causal antecedent (cause) and a causal consequence (effect). This is different from the Process predicate type as it has one aspect that governs another aspect even if it is not explicitly stated. This often takes the form of direct or inverse mathematical relationships.

a) Tutor: What is the connection between the distance between the magnets and the strength of the force?

(Ask/Describe/CausalRelation)

Student: As the magnets get closer, the strength gets bigger

(Answer/Describe/CausalRelation)

A.3.0.8 Configuration

Describes a configuration or setup.

a) Tutor: What was the orientation of the poles?

(Ask/Describe/Configuration)

Student: The opposite poles are facing each other.

(Answer/Describe/Configuration)

A.3.0.9 Duration

Used when referencing a length or span of time.
a) Tutor: *How long does the metal stay magnetized?*  
(Ask/Describe/Duration)  

Student: *It stays magnetized for a few seconds*  
(Answer/Describe/CDuration)

**A.3.0.10 Entity**

Used when referring to objects or entities. This is often used within visual contexts.

a) Tutor: *Which one is bigger?*  
(Ask/Compare/Entity)  

Student: *The one on the right is bigger.*  
(Answer/Compare/Entity)

**A.3.0.11 Experience**

Used to mark that an utterance references past experience or knowledge.

a) Tutor: *Remember what it felt like to try and push together two magnets that didn't want to stick. What was that all about?*  
(Ask/Recall/Experience)

**A.3.0.12 Function**

Used to describe a purpose, function or usage of some referent.

a) Tutor: *Tell me more about what a lightbulb does.*  
(Ask/Describe/Function)  

Student: *It takes electricity and makes light from it.*  
(Answer/Describe/Function)

**A.3.0.13 Location**

Used to denote a location, but not a trajectory like Route does.
a) **Tutor:** Where do you connect the wires?

(Ask/Describe/Location)

**Student:** You hook them up to the contact points.

(Answer/Describe/Location)

---

A.3.0.14 **Observation**

Used to describe something observed visually. This is different from Ask/Describe/Visual which is a more literal description. An utterance with an Observation tag should be describing some relation or process observed.

a) **Tutor:** What happens when you click on the magnet?

(Ask/Describe/Observation)

**Student:** It sticks to metal things.

(Answer/Describe/Observation)

---

A.3.0.15 **Procedure**

Used to describe a procedure or a sequence of steps/actions.

a) **Tutor:** Let’s talk more about your work with washers, magnets and the balance. Tell me more about what you did

(Ask/Describe/Procedure)

**Student:** we dropped washers into a cup and there were two magnets there sticking together and once there’s a certain amount of washers then the magnets wouldn’t come apart and then we put a little spacers that were almost like little poker chips in between the magnets and the more of those that you put the less washers it would hold

(Answer/Describe/Procedure)
A.3.0.16 Process

Describes a phenomena or process. A natural or involuntary event or series of events that causes a change in state this is often used to describe the occurrences in a visual.

a) Tutor: *I think I heard you say that magnets stick or attract. Tell me more about that.*

   (Revoice/-/-, Ask/Elaborate/Process)

Student: *Magnets attract to metal.*

   (Answer/Elaborate/Process)

A.3.0.17 Proposition

Used to denote that an utterance refers to a target proposition (concept).

a) Tutor: *Electricity flows from the positive end of the battery to the negative end.*

   (Assert/Recap/Proposition)

A.3.0.18 Route

Used to describe a path, which consists of a beginning and ending point and possibly the medium it goes through.

a) Tutor: *Tell me more about the direction electricity flows from the d cell.*

   (Ask/Describe/Route)

Student: *It goes from positive to negative*

   (Answer/Describe/Route)

A.3.0.19 Topic

Used to refer to some topic or subject matter.
a) **Tutor:** Tell me about measurement.

(Ask/Describe/Topic)

**Student:** Measurement is used for knowing the size of stuff

(Answer/Describe/Topic)

A.3.0.20 Visual

Pertaining to a visual or other form of interactive multimedia

a) **Tutor:** What’s going on in this picture?

(Ask/Describe/Visual)

**Student:** There’s a ball and a nail and other stuff

(Answer/Describe/Route)

A.3.0.21 YesNoMaybe

Used with rhetorical form YesNo to denote the yes/no response.

a) **Tutor:** Is the circuit on now?

(Ask/YesNo/-)

**Student:** Nope.

(Answer/YesNo/YesNoMaybe)
Appendix B

DISCUSS Classifiers Error Analysis

This appendix provides detail on the errors for each of the DISCUSS dimensions and labels. Since it is not practical to list all of the errors for all of the DISCUSS labels, the focus will be on highlighting common sources of mistakes for each of the classifiers.

Two phrases used heavily throughout the remainder of this appendix are “False Positive” and “False Negative”. These terms refer to two types of mis-classifications made by the DISCUSS utterance labeler. A “False Positive” is an error of commission. That is falsely assigning a label to utterance. For example if an utterance was annotated with the tuple \textit{Answer/Identify/Entity} tagging it with a rhetorical form of \textit{Describe} would be a false positive for the \textit{Describe} classifier. Similarly, “False Negative” refers to the classifier’s omission of a gold-standard tag. Given the DISCUSS tuple described above, a false negative would arise from not labeling the utterance with rhetorical form \textit{Identify}.

B.1 Dialogue Acts

B.1.1 Acknowledge

\textbf{False Positives}: As a grounding act, acknowledgment can often look like agreement or confirmation. The majority of the false positive errors were from short 1-2 word utterances that in different contexts would look more like an acknowledge.
a) **Student:** okay  
b) **Student:** well  
c) **Student:** yeah  
d) **Student:** mm hmm yes

### B.1.2 Answer

The *Answer* tag scored had the best performance of all the Dialogue Act classifiers, most likely because this was the majority class with over 80% of all student utterances exhibiting the presence of this tag. Unlike with the other DISCUSS tags, the challenge for *Answer* lies in minimizing the false positive rate.

**False Positives:** Several false positives arose from annotator mistakes. In the following instance, the annotator had assumed the student was not answering the question and labeled the expression as *Assert/Describe/Procedure*.

a) **Tutor:** Very good. Go ahead and click through this again. What did adding more washers do to the force between the magnets?  
**Student:** add more washers it

Questions asked by students posed a challenge for the *Answer* classifier as there are lexically few cues to differentiate it from an *Ask*. Features more targeted at identifying question surface forms could help to eliminate these errors.

b) **Student:** how can centimeters how many centimeters are in a yard

Metastatements uniformly caused difficulty as they do not exhibit the same regularity as other classes.

c) **Student:** no i i wasn’t pressing the spacebar that whole time

**False Negatives:** One word answers provide little evidence for classifying answer.
d) Student: yes

e) Student: no

f) Student: magnet

B.1.3 Close

**False Positives:** Expression of thanks are often spoken in conjunction with closing and saying goodbye. In some cases, the annotators even marked “thank you” and “you’re welcome” as Close. With limited training data, the following utterances can be mistaken for Close acts.

   a) Student: thank you

   b) Student: you’re welcome

B.1.4 Open

**False Positives:** In many Open utterances, the student answers “Good” in response to “How are you doing today?”. The following error suggests a possible overfitting to the presence of this word.

   a) Student: good bye

**False Negatives:** The name “Marnie” (the name of the tutor within MyST) was predominantly spoken at the tail end of the session when the student was saying goodbye. Consequently the presence of “Marnie” in a short utterance confused the classifier, even though the student was using a straightforward “Hello”. This could be fixed with additional training data, or more realistically, with hand crafted rules or grammars.

   b) Student: hello marnie

B.1.5 Thank

**False Negatives:** Two out of five of the Thank false negative errors included the word “thanks”. Closer inspection of the lemma based features found that the morphological analyzer (lemmatizer) did not truncate “thanks” to “thank”.

B.1.6 Metastatement

58% of the falsely identified Metastatement utterances included the either the word “yes” or “yeah”. While these words are high runners within answers to yes/no questions, they also show up in 14% of all Metastatements in the corpus. In many cases these are incorrectly labeled by the annotator. The rarity and inconsistency in metastatements does not leave much room for improvement for this class, and future effort would be better spent ignoring this class.

B.1.7 SignalNoUnderstanding

False Positives: From a tutor’s perspective Uninterpretable and SignalNoUnderstanding are nearly identical, and both signal that the student is struggling with the material. The false positives, reinforce this notion with many utterances that were marked Uninterpretable are really truncated or cut-off versions of SignalNoUnderstanding.

a) Student: i don’t quite
b) Student: i don’t

False Negatives: Unseen expressions that effectively mean “I don’t know:” went undetected by the classifier.

c) Student: i can’t think of anything
d) Student: that’s kinda a tough one for me
e) Student: i forgot

B.1.8 Uninterpretable

False Positives: For the same reasons as stated above SignalNoUnderstanding statements were often confused with Uninterpretable statements.

False Negatives: The phrases and utterances flagged negative by the classifier exhibit very little regularity in words or speech patterns aside from a lack of domain specific words.
B.1.9 MergedNoUnderstanding

The error analyses for SignalNoUnderstanding and Uninterpretable showed consistent confusion between the two classes and the small number of overall examples gave strong motivation to merge these two labels. This merging greatly decreased the false positive and false negative rates for each of these classes respectively.

**False Positives:** The presence of negation words such as ”not”, “don’t” and “know” are strong indicators of MergedNoUnderstanding, and are consequently difficult to override. Adding features that capture the presence of domain specific words may prevent this error.

a) Student: because you don’t know what’s in it there might been i don’t know something like cheese

**False Negatives:** Uninterpretable utterances are often composed entirely of frequently used words and disfluencies. In some cases they also appear interrupted. Features derived from stopword lists or TF*IDF scores can possibly address this issue of lexical diversity. Additionally, prosodic features from the audio stream may provide more reliable predictive capabilities.

b) Student: it is like a like a
c) Student: the the
d) Student: so it so the so the

B.2 Rhetorical Forms

B.2.1 Bye

The errors for Bye are nearly identical to those made by the Close dialogue act classifier.
B.2.2 Compare

**False Positives:** The errors for *Compare* illustrate the advantages of allowing multiple tags per utterance rather than forcing a strict one to one matching. The three utterances below were classified as *Compare*. Even though this rhetorical form differs from the gold-standard rhetorical forms chosen by the annotators, one could argue that these utterances exhibit properties from both forms and are indeed correct.

a) **Student:** *there are two circuits one on the right and one on the left*  
   (Answer/Quantify/Entity)

b) **Student:** *and the circuit on the on the left side is the light bulb is brighter*  
   *because there’s only one light bulb and on the on the right side*  
   *there’s two light bulbs so it’s*  
   (Answer/Describe/Observation)

c) **Student:** *some of the straws were longer than the other ones and some of*  
   *them were shorter than the other*  
   (Answer/Elaborate/Attribute)

**False Negatives:** For some utterances, the comparison is implicit and can only be inferred from context. Though the previous turn features do capture this context, the lack of firm lexical cues like comparatives and quantifiers are the likely cause for the following misclassification of the student utterance.

d) **Tutor:** *This is a parallel circuit. How does this compare to other circuits you have made?*

   **Student:** *the wires don’t touch each other and they and there’s two wires connected to the positive and negative side unlike some other times there’s only one connected to the positive and one connected to the negative*
B.2.3 Confirm, YesNo and MergedYesNo

From a surface form perspective, student responses Confirm and YesNo are indistinguishable from one another, and in terms of information content they serve similar roles. Though the classifiers did not confuse these two labels, they were prone to errors of omission when classifying lexically similar utterances with different dialogue acts such as Acknowledge. The challenge then shifts to determining when a 'yes' response is an answer to a yes or no question versus acknowledging what was said.

Combining Confirm and YesNo reduces ambiguity and shifts the focus toward identifying these specific types of “Yes” or “No” responses. The combined MergedYesNo class reduced the total number of false negatives and increased $F_1$-score.

**False Negatives:** Coincidentally, negatively phrased responses were the source of many of the false negatives. These errors most likely stem from sparsity in the training data and would be best addressed by hand crafting grammars, rules or new features to better address specific phrases.

a) Tutor: does schematic sound familiar?  
Student: no it doesn’t

B.2.4 Define

**False Positives:** Differentiating between Define and Describe was difficult for the human annotators, and is reflected in the collection of utterances falsely identified as having Define rhetorical forms. The following examples show false positives that could be considered to exhibit the characteristics of a Define rhetorical form.

a) Student: length is like is like from is like how long it is from one place to a whole nother place  
b) Student: measurement is when is when you see like how tall something is how wide something is or how long something is  
c) Student: measuring is about like giving an answer on how long or how wide something is or even how tall
**False Negatives:** Like with false positives, the ambiguity between *Define* and *Describe* factored into the misclassifications. In these instances, it is unclear where the boundary between the two classes lie.

   d) **Student:** measuring is about like giving an answer on how long or how wide something is or even how tall

   e) **Student:** it’s the source of energy

   f) **Student:** a closed a closed circuit will be able to have electricity flow through and or a and the circuit will be completed and the receiver will be able to receive the electricity

**B.2.5 Describe**

**False Positives:** As the default class, *Describe* is prone to confusion with other rhetorical forms that are closely coupled to the tutorial subject domain, especially when utterances are long. The following four errors highlight ambiguity in the rhetorical form. The most common confusion occurs with the *Elaborate* tag.

   a) **Student:** the screen sticks to the magnet the rusty nail sticks to the magnet the ring sticks to the magnet the pen sticks to the magnet and the paper clip sticks to the magnet that’s what sticks to magnet

   (Answer/List/Entity)

   b) **Student:** it’s flowing the electricity is flowing out of the negative side and into the positive side of the battery

   (Answer/Identify/Location)

   c) **Student:** the wires are they are it’s what the electricity travels through

   (Answer/Elaborate/Function)

   d) **Student:** repel means that they’re not attracted to to each other

   (Answer/Define/CausalRelation)
**False Negatives:** A majority of the false negatives come from short utterances without many lesson keywords. The confusion with the *Elaborate* rhetorical form also plays a role. In the following cases, the preceding tutor utterance is labeled with *Elaborate*.

e) **Tutor:** *Tell me more about the objects that are insulators.*  
(Ask/Elaborate/Entity)  
**Student:** *they don’t let metal flow through them they’re just kinda solid for with electricity*

**B.2.6 Elaborate**

**False Positives:** Of the false positives approximately 80% are tagged with *Describe* for the rhetorical form. Again, this makes a strong case for merging classes to get more consistent results.

**False Negatives:** Like with the false positives, there are not many cues to distinguish *Elaborate* from *Describe* save for the rhetorical form dictated by the previous question. Given the size of the feature space, the importance of this feature is down-weighted as an *Answer/Elaborate* is just as likely to follow *Ask/Elaborate* as an *Answer/Describe*.

**B.2.7 Greet**

The errors made by the *Greet* rhetorical form classifier parallel those made by the *Open* dialogue act, which suggests that this label is most redundant.

**B.2.8 Identify**

**False Positives:** Identifying utterances are often characterized by the use of copula, especially of the form “X is Y”. In some cases this helps to discover utterances that exhibit *Identify*-like qualities even when the annotator did not label it that way.
a) **Student:** the twirl with the circle over it is a light bulb and the two ts are the battery and the lines are the wires

b) **Student:** that is a different kind of parallel circuit

c) **Student:** the middle’s a battery and the sides are the light bulb and the lines are the wires

On the other hand, copula alone do not predict *Identify*. While the classifier can learn copula patterns from the POS/word unigrams and bigrams. These n-gram sizes are not long enough to learn the full sequences associated with *Identify*. The small size of the WOZ corpus prevents scaling to larger n-grams. Preventing errors like those shown below would instead require more explicit patterns that could be applied during feature extraction, pre-processing or post-processing. The pattern [noun] [copula] [determiner] [noun] would eliminate the following errors. More sophisticated patterns may require the use of a syntactic parser or noun chunking.

d) **Student:** in my own words magnets are attracted to iron

e) **Student:** it’s a lot dimmer

f) **Student:** i think it’s ten inches

### B.2.9 Justify

**False Positives:** Beyond the presence of *Justify* in the preceding turn’s rhetorical form, there are few consistent cues for classifying this label. The word “because” is one of the few trigger words for *Justify*, but taken alone it can easily confuse descriptions of causal relations with a student’s reflection on his or her reasoning and thought processes.

a) **Student:** because the aluminum nail does not have iron in it

b) **Student:** it tells me that i found it because when i tried to place it somewhere it snuck over to a different place and it stuck there so

c) **Student:** everybody would get a different answer because they different lengths so nobody would get the exact answer

**False Negatives:** Within the WOZ corpus, there are 254 utterances that contain the word “because”. Of these only 5.5% are labeled *Justify*. Thus, even if “because” is an indicator of
justification, it is not a strong enough indicator to override other evidence used by the classifier. Consequently, nearly half of the false negatives contain the word “because”.

d) **Student:**  *well i knew there’s two ways because we did it in class and we we all got to and because if the iron sticks to the spot on the box you know there’s something like a magnet in*

e) **Student:**  *it made me think that because iron magnets help you magnets help you find things that’s why i thought about that cause like that*

**B.2.10 List**

**False Positives:** A canonical *List* rhetorical form utterance would consist of a list of nouns separated by commas and a conjunction. With this spoken data, there is no punctuation, consequently the primary signal becomes sequencing of nouns and coordinating conjuncts. The false positive rate could be decreased by using more sophisticated syntactic patterns to identify list sequences.

a) **Student:**  *well iron ; iron can conduct the electricity and stick and also stick to magnets*

b) **Student:**  *twelve and eighteen*

c) **Student:**  *the north and south attract and north and north or south and south repel*

**False Negatives:** Inconsistency in annotation is the largest barrier to accurate classification of the *List* rhetorical form. The low inter-annotator agreement of $\kappa = 0.31$ for *List* reflects the difficulty annotators had in determining what constituted a list. In particular they struggled with whether or not a response to a question with DISCUSS tags *Ask/List/* was automatically a list. Combining *List* with non-*Entity* predicate types, further confounded the annotators. The errors below reflect this difficulty and highlights some of the ambiguity associated with the *List* tag.
d) Student: in my own words magnets are attracted to iron
   (Answer/List/Entity)

e) Student: the insulators do not conduct electricity unlike the conductors conduct electricity
   (Answer/List/Attribute)

f) Student: if you need help you can ask somebody or you can try to use a measurement tool
   (Answer/List/Procedure)

B.2.11 Quantify

**False Positives:** Numeric words are an important feature for identifying the Quantify rhetorical form. At the same time, numeric words are used in a variety of ways, and may not necessarily translate to a quantification. The errors below illustrate such instances.

a) Student: there is two pathways in the circuit

b) Student: this is a schematic diagram there are two batteries and two light bulbs

c) Student: well in one centimeter there are ten millimeters

**False Negatives:** As with most DISCUSS labels, the most likely cause for errors in recall stem from sparsity in training data. While number normalization could improve these errors, it would need to be done carefully so as not to break numbers used in different contexts. Furthermore, the potential gains are likely limited as the classifier already includes a numeric word presence feature.

B.2.12 MergedDescribe

Merging the Define, Describe and Elaborate greatly improved performance for each of these three classes with number of false positive and false negative instances falling to a third of the original. As the majority class, the presence of MergedDescribe in a student response provides nearly no information gain when deciding how to ask follow-up questions. This label can be viewed as the default, and it is the absence of this label that carries more discriminative power.
False Positives: Utterances labeled *Compare*, *List* and *Identify* continue to confound the merged classifier. This is expected behavior as these constitute the bulk of the rhetorical form instances after the merged *MergedDescribe* class. The patterns and rules suggested in the section for each its constituent rhetorical forms would apply to classifying this merged class as well.

a) **Student:** *and it’s iron or steel*  
(Answer/Identify/Attribute)

b) **Student:** *well i see paperclips wire lookings thing a rock and a sponge and nail some sort of glass tube and a tongue depressor and a brass ring looks like a brown rock and a black rock*  
(Answer/List/Observation)

c) **Student:** *one is broken and the other one is perfect*  
(Answer/Compare/Entity)

False Negatives: The false negatives are a mixture of several kinds of errors that do not appear to have any singular cause. Among the many types of errors, the *MergedDescribe* classifier struggled the most with short student responses. Utterances five or less words in length comprised only 6.6% of all positive instances (correct and incorrect). Conversely, they only made up 56% of the instances falsely classified as having no *MergedDescribe* label. The small number of words give little evidence to allow a positive classification.

d) **Student:** *three feet*

e) **Student:** *magnets*

f) **Student:** *mostly it’s metal*

B.2.13 MergedYesNo

Combining *YesNo* and *Confirm* into one class removed the lexical confusion that could confuse or weaken the classifiers’ predictions.

False Positives: All of the false positives came from one to two word “yes” responses whose function was not an answer to a yes or no question, but as an acknowledgment or grounding.
a) Tutor:  
*great job, series circuits share the electricity*  
(Feedback/Positive/None)

Student: *mm hmm yes*  
(Acknowledge/None/None)

b) Tutor:  
*So we have observed that magnetism is a force that goes through most materials and that the force of magnetism decreases as distance increases*  
(Assert/Recap/Proposition)

Student: *yes*  
(Acknowledge/None/None)

**False Negatives:** In contrast to the false positives, the false negatives mainly occurred when the “yes”s or “no”s were paired with other descriptions.

a) Student: *three feet*

b) Student: *magnets*

c) Student: *mostly it’s metal*

**B.3 Predicate Types**

A common theme in the error analysis of predicate type classification is dependence on lesson domain. While dialogue context and syntactic construction may help in predicting the predicate type, the dominant features are the key vocabulary from the FOSS subject domains. Another common theme is the multifaceted nature of spoken dialogue. Often a single predicate type can not fully express the content of a tutor’s question and student’s answer. Thus the errors may be more of an artifact of the annotation than a definitive evaluation of incorrectness.

**B.3.1 AcceptRejectMaybe, YesNoMaybe, and MergedYesNoMaybe**

The need to merge *AcceptRejectMaybe* and *YesNoMaybe* predicate types arises from the same motivations for combining the *Confirm* and *YesNo* rhetorical forms. At the heart of the issue is
strong lexical overlap the confuses the classifiers.

Like the lexicon, the errors follow directly from their companion rhetorical forms. This suggests that without the finer grained ability to differentiate Yes's, No's, and Maybe's, there is no additional information to be gained from these predicate types.

**B.3.2 Activity, Experience, Topic, and MergedActivity**

Though Activity, Experience and Topic may not seem likely candidates for merging into one class. The error analyses show that there is large overlap in how the annotators applied these tags. This chain of ambiguity, poses difficulty for the classifier learning algorithms, and merging is the best way to eliminate this noise.

**False Positives:** Nearly seventy percent of the false positive errors for Topic were from utterances tagged with predicate type Activity, over fifty percent of falsely identified Activity tags were for Topic utterances. Examples (a)-(d) demonstrate the ambiguity between these two predicate types. While (a) and (b) were incorrectly tagged as Topic instead of Activity, examples (c) and (d) where from the opposite condition.

a) **Student:** we’ve been doing parallel circuits

b) **Student:** we’ve been experimenting with circuits with batteries wires and motors or lights and switch

c) **Student:** we’ve been studying magnetism about magnets

d) **Student:** we’ve been learning about measurements

Because there were few Experience-tagged utterances, the errors for this classifier were not as numerous or reciprocal as Activity and Topic. Still, a large percentage of Experience false positive errors are representative of this overlap.

e) **Student:** we’ve been measuring our desk with straws

f) **Student:** in class i we were doing we had these balancing things to to balance the magnet
False Negatives: As with many of the less frequent DISCUSS predicate types, the small number of positive training examples for Experience hurt recall. Collisions with Activity further compounded false these issues. This ambiguity effectively eliminated the effect of positively labeled Experience utterances during training.

There were two categories of Topic false negative errors 1) abbreviated utterances and 2) definition answers. The one-word student response in example (a) shows a typical abbreviated response with predicate type Topic. One possible solution for this would be to integrate additional contextual features. In definition responses, like those listed in examples (b) and (c), were difficult to classify because their features did not share many similarities with the more Activity-like utterances, and instead closely resembled utterances with different predicate types. Future annotation effort could reduce ambiguity by creating a new Term predicate type.

a) Tutor: What have you been doing in science class?
   Student: magnets

b) Student: parallel circuits are when they the wires don’t don’t touch each other in their circuit and have separate circuits for each light bulb

c) Student: magnetism is what comes out of a magnet

Merging these three predicate types into the MergedActivity label eliminated many of these issues and produces a 0.1 gain in overall F₁-score.

B.3.3 Attribute

False Positives: The Attribute classifier was prone to overly optimistic classification because of the strong association between adjectives and attributes/properties. Furthermore the small corpus size tends toward overfitting to very lesson-specific features such as the presence of words including “iron”, “steel”, or “aluminum”, or even more general words such as “two”.
a) **Student:** well it has two pathways
   \[\text{Answer/Describe/Configuration}\]

b) **Student:** and it connects to anything that has iron in it
   \[\text{Answer/Describe/CausalRelation}\]

c) **Student:** it didn’t pick the aluminum nail
   \[\text{Answer/Describe/Observation}\]

### B.3.4 CausalRelation, Process and MergedCausalRelation

As the majority predicate type class *CausalRelation* exhibits poor precision because the classifiers are too optimistic. Conversely, *Process* classification performance suffers from sparsity of positive training examples. At the same time, annotators struggled to differentiate between the two predicate types. The results show merging these classes gives a large boost in eliminating the mistakes made for *Process*. Examples (a)-(d) demonstrate the challenge in disambiguating *CausalRelation* and *Process*.

**False Positives:**

a) **Student:** well the circuit breaks because the electrons go through that burned out part
   \[\text{Answer/Describe/Process}\]

b) **Student:** because the aluminum nail does not have iron in it
   \[\text{Answer/Describe/Process}\]

c) **Student:** magnetic force can go through cardboard plastic and and paper
   \[\text{Answer/Describe/CausalRelation}\]

d) **Student:** aluminium doesn’t attract to magnets
   \[\text{Answer/Describe/CausalRelation}\]

### B.3.5 Configuration

**False Positives:** The semantics of *Configuration* is tied closely tied to specific subject matter. In the domain of circuits and electricity, the phrases “parallel circuit” and “series circuit” describe
specific configurations. Thus their presence often biased the system towards a positive classification. In some cases the confusion came down to differences in annotator perception. In the following error, the annotator considered the question and answers to revolve around an Entity; however it can just as likely be considered a question about configuration.

a) Tutor:  *Tell me about the kind of circuit you see here.*

(Ask/Identify/Entity)

Student:  *a parallel circuit*  

(Answer/Identify/Entity)

**False Negatives:** The false negatives illustrate the ambiguity in annotating predicate types. Consider the question-answer pair below. The tutor’s question in example (b) is asking the student to describe a causal relation. The student’s response could be interpreted in two ways. The annotator’s literal interpretation yielded a tag of Answer/Describe/Configuration. A more contextual annotation would have labeled the same utterance as Answer/Describe/CausalRelation.

b) Tutor:  *How does what is happening with the compass help you figure out what the compass is made from?*

Student:  *there’s a small magnet in the compass*

**False Positives:** The Entity classifier depends on the presence of domain specific keywords to make its decisions. If the feature vector for a student utterance includes word features like “wire”, “light bulb” or “battery”, it is more likely to be about an entity. At the same time, these words can be used into more complex predicate types such as describing functions or causal relations. This duality leaves Entity classification prone to overly-optimistic labeling like with the utterances below.
a) **Student:** the d cell is a source of power

(Answer/Describe/Function)

b) **Student:** circuits are about making things work like a motor

(Answer/Describe/Function)

c) **Student:** the magnet is in the box right there

(Answer/Identify/Location)

**B.3.7 Function**

**False Positives:** Discussion of an objects function often leads to discussion of cause and effect, and so utterances which ultimately describe a *CausalRelation* can be confused with *Function*. In this example, the tutor is asking for a function, and the student automatically bridges to the causal relation. A more correct gold-standard annotation would have included both labels.

a) **Tutor:** Tell me more about what the wires do in a circuit

(Ask/Describe/Function)

**Student:** they bring the energy from the battery to the motor which powers the motor

(Answer/Describe/CausalRelation)

**False Negatives:** Many of the false negative classifications were caused by annotator error. The sequence below really centers on asking about an *Attribute* or *Configuration* of the pathway, not its *Function*.

b) **Tutor:** Cool, so what is important about the pathway of electricity in the series circuit?

**Student:** the electricity has to be shared within the two receivers

**B.3.8 Location, Route and MergedRoute**

Semantically *Location* could be considered one dimension of a *Route*. From a surface-form perspective, utterances from these tags share many of the same lexical and syntactic features. This
similarity causes confusion for both machine classifiers and human annotators and argue for merging these tags into a single *MergedRoute* tag.

**False Positives:** Prepositions are an important feature for identifying both the *Location* and *Route* predicate types. Example (a) is a *Route* utterance misclassified as a *Location* utterance, and example (b) is the opposite condition.

a) **Student:** *it flows out of the bottom of the d cell and goes into the bottom*  
(Answer/Describe/Route)

b) **Student:** *minus side the electricity from the d cell is flowing through the minus side to the light bulbs and coming in through the plus side*  
(Answer/Identify/Location)

**False Negatives:** The similarity between *Location* and *Route* was again a factor in false negative errors. When learning to classify *Location* the large number of *Route* utterances became negative examples, which in turn discounted features that should assist in identify *Location* predicate types.

c) **Student:** *electricity is going out of the d cell battery and lighting the light*

**B.3.9 Observation, Visual and MergedVisual**

In the annotation guidelines, *Observation* was intended for higher-level descriptions of on-screen activity, whereas *Visual* was supposed to be for more concrete language. In practice, the annotators would use these tags interchangeably. Like with other highly overlapping tags, *Observation* and *Visual* were merged into *MergedVisual* to make a more robust predicate type classifier.

**False Positives:** Aside from being confused with one another, *Observation* and *Visual* the most frequent error of commission occurred when labeling *CausalRelation* utterances. This error is not unexpected, since *CausalRelation* is the majority predicate type class and because the boundary between describing what is on screen and describing a cause and effect sequence is underspecified. Consider example (a). Though this was labeled with predicate type *CausalRelation*, an equally compelling case can be made for *Visual*. 

a) Student: there is one light bulb broken so so the other light bulb doesn’t turn on

B.3.10 Procedure

False Positives: With MyST dialogues procedural language is often spoken in the first or second person with high co-occurrence with verbs such as “put”, “have”, and “make”. At the same time causal relations can be described in a procedural style, especially when referring to objects on screen. The false positives errors largely followed these patterns, but did not have Procedure as its annotated predicate type.

a) Student: don’t put any spaces in you get a higher number but if you put five spaces in you get like three or two or one or something

(Answer/Describe/Observation)

b) Student: if you have more spacers the magnets force will be less so therefore your washers that you put in will turn out less because you have the spaces in between the magnets

(Answer/Describe/CausalRelation)

False Negatives: In contrast to the false positive errors, the lack of first or second person could cause a negative classification.

c) Student: follow the instructions and be creative
d) Student: added more washers it

Alternatively, very long utterances were often mistaken as other predicate types:

e) Student: you attach the wires to each side so that the positive energy goes out one side and then the negative energy comes out from the bulb and goes back into its side
Appendix C

Correlation with Learning Gains Features

C.1 Basic Features

**AVERAGE_ONSET_TIME_STUDENT** The average time, in seconds, between the end of a tutor’s turn and the beginning of a student’s turn

**AVERAGE_ONSET_TIME_TUTOR** The average time, in seconds, between the end of a student’s turn and the beginning of a tutor’s turn

**AVERAGE_TURN_TIME_TUTOR** The average time, in seconds, of a tutor’s turn

**AVERAGE_TURN_TIME_STUDENT** The average time, in seconds, of a student’s turn

**AVERAGE_WORDS_PER_TURN_STUDENT** The average number of words per student turn

**AVERAGE_WORDS_PER_TURN_TUTOR** The average number of words per tutor turn

**PERCENT_TIME_STUDENT** The percent of the total dialogue turn time spoken by the student

**PERCENT_TURNS_STUDENT** The percent of the total number of turns spoken by the student

**PERCENT_WORDS_STUDENT** The percent of the total words in a dialogue spoken by the student
C.2 Phoenix Dialogue Manager Features

AVERAGE_TIMES_PROMPTED_PER SLOT Measures the degree of prompting needed to elicit the target student speech. The higher this value, the more on average the tutor prompted on a specific point to elicit a response.

PERCENT_PROMPTS_FROM_RULES The percent of tutor moves (i.e. prompts) that were triggered by rules

PERCENT_SLOTS_FAILED The percent of slots (frame-elements) in the task file that were prompted for, but left unfilled at the end of the tutoring session(s)

PERCENT_STUDENT_TURNS_PARSE_INTO_CURRENT_FRAME The percentage of turns spoken by the student that were successfully parsed by Phoenix that also filled slots within the current dialogue frame

PERCENT_STUDENT_TURNS_UNPARSEABLE The percentage of turns spoken by the student that Phoenix was unable to parse.

C.3 DISCUSS Features

AVG_DISTANCE_TO_MINIMAL_FRAME_SEQUENCE In this feature distance refers to the edit distance between the observed sequence of DISCUSS tuples through a Phoenix frame and the shortest possible sequence to frame completion. This is averaged across all frames and dialogues for a given student.

STUDENT_TURNS_DA_{\langle TAG\rangle}\% The percentage of student turns in the dialogue(s) that have the Dialogue Act (DA) specified by \langle TAG\rangle (e.g. Ask, Assert, Revoice, etc...)

TUTOR_TURNS_DA_{\langle TAG\rangle}\% The percentage of tutor turns in the dialogue(s) that have the Dialogue Act (DA) specified by \langle TAG\rangle (e.g. Ask, Assert, Revoice, etc...).
STUDENT_TURNS_RF_{TAG}\% The percentage of student turns in the dialogue(s) that have
the Rhetorical Form (RF) specified by \langle TAG \rangle (e.g. MergedDescribe, Compare, Quantify,
etc.)

TUTOR_TURNS_RF_{TAG}\% The percentage of tutor turns in the dialogue(s) that have the
Rhetorical Form (RF) specified by \langle TAG \rangle (e.g. MergedDescribe, Compare, Quantify, etc.)

STUDENT_TURNS_PT_{TAG}\% The percentage of student turns in the dialogue(s) that have
the Predicate Type (PT) specified by \langle TAG \rangle (e.g. MergedVisual, Procedure, Function,
etc.)

TUTOR_TURNS_PT_{TAG}\% The percentage of tutor turns in the dialogue(s) that have the
Predicate Type (PT) specified by \langle TAG \rangle (e.g. MergedVisual, Procedure, Function, etc.)

TUTOR_TURNS_DARF_PT_{TUPLE}\% The percentage of tutor turns in the dialogue(s) that have
the DISCUSS tuple specified by \langle TUPLE \rangle (e.g. Ask_MergedDescribe_Function)

PERCENT_TURNS_PT_MATCH The percent of turns whose Predicate Type (PT) matches
the predicate type of the previous turn

PERCENT_TURNS_RF_MATCH The percent of turns whose Rhetorical Form (RF) matches
the predicate type of the previous turn