Question Ranking and Selection in Tutorial Dialogues

Lee Becker[a] and **Martha Palmer 2**[a] and **Sarel van Vuuren 3**[a] and **Wayne Ward 4**[a,b]
[a]The Center for Computational Language and Education Research (CLEAR)
University of Colorado Boulder
[b]Boulder Language Technologies
{lee.becker,martha.palmer,sarel.vanvuuren}@colorado.edu
wward@bltek.com

Abstract

A key challenge for dialogue-based intelligent tutoring systems lies in selecting follow-up questions that are not only context relevant but also encourage self-expression and stimulate learning. This paper presents an approach to ranking candidate questions for a given dialogue context and introduces an evaluation framework for this task. We learn to rank using judgments collected from expert human tutors, and we show that adding features derived from a rich, multi-layer dialogue act representation improves system performance over baseline lexical and syntactic features to a level in agreement with the judges. The experimental results highlight the important factors in modeling the questioning process. This work provides a framework for future work in automatic question generation and it represents a step toward the larger goal of directly learning tutorial dialogue policies directly from human examples.

1 Introduction

Socratic tutoring styles place an emphasis on eliciting information from the learner to help them build their own connections to the material. The role of a tutor in a Socratic dialogue is to scaffold the material and present questions that ultimately lead the student to an "A-ha!" moment. Numerous studies have illustrated the effectiveness of Socratic-style tutoring (VanLehn et al., 2007; Rose et al., 2001; Collins and Stevens, 1982); consequently recreating the behavior on a computer has long been a goal of research in Intelligent Tutoring Systems (ITS). Recent successes have shown the efficacy of conversational ITS (Graesser et al., 2005; Litman and Silliman, 2004; Ward et al., 2011b), however these systems are still not as effective as human tutors, and much improvement is needed before they can truly claim to be Socratic. Furthermore, development and tuning of tutorial dialogue behavior requires significant human effort.

While our overarching goal is to improve ITS by automatically learning tutorial dialogue strategies directly from expert tutor behavior, we focus on the crucial subtask of selecting follow-up questions. 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.

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, we do not restrict ourselves only to the moves and questions found in a corpus of transcripts. Instead we posit "What if we had a fully automatic question generation system?" and subsequently use candidate questions hand-authored for each dialogue context. To explore the mechanisms involved in ranking follow-up questions against one other, we pair these questions with judgments of quality from expert human tutors and extract surface form and dialogue-based features to train machine learning classification models to rank the appropriateness of questions for specific points in a dialogue.

Our results show promise with our best question

The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 1–11,
Montréal, Canada, June 3-8, 2012.

ranking models exhibiting performance on par with expert human tutors. Furthermore these experiments demonstrate the utility and importance of rich dialogue move annotation for modeling decision making in conversation and tutoring.

2 Background and Related Works

Learning tutorial dialogue policies from corpora is a growing area of research in natural language processing and intelligent tutoring systems. Past studies have made use of hidden Markov models (Boyer et al., 2009a) and reinforcement learning (Chi et al., 2010; Chi et al., 2009; Chi et al., 2008) to discover tutoring strategies. However, these approaches are typically optimized to maximize learning gains, and are not necessarily focused on replicating human tutor behavior. Other work has explored specific factors in questioning such as when to ask "why" questions (Rose et al., 2003), provide hints (Tsovaltzi and Matheson, 2001), or insert discourse markers (Kim et al., 2000).

There is also an expanding body of work that applies ranking algorithms toward the task of question generation (QG) using approaches such as overgeneration-and-ranking (Heilman and Smith, 2010), 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, we move to the higher level task of choosing context appropriate questions. Our work 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, we use them directly as part of the feature space for learning functions to rank the questions' actual surface form realization. To our knowledge there has been no research in ranking the quality and suitability of questions within a tutorial dialogue context.

Because questioning tactics depend heavily on the curriculum and choice of pedagogy, we ground our investigations within the context of the My Science Tutor (MyST) intelligent tutoring system (Ward et al., 2011b), a conversational virtual tutor designed to improve science learning and understanding for students in grades 3-5 (ages 8-11). Students using MyST investigate and discuss science through natural spoken dialogues and multimedia interactions with a virtual tutor named Marni. The MyST dialogue design and tutoring style is based on a pedagogy called Questioning the Author (QtA) (Beck et al., 1996) which emphasizes open-ended questions and keying in on student language to promote self-explanation of concepts, and its curriculum is based on the Full Option Science System (FOSS) [1] a proven system for inquiry based learning.

3 Data Collection

3.1 MyST Logfiles and Transcripts

For these experiments, we use MyST transcripts collected in a Wizard-of-Oz (WoZ) condition with a human tutor inserted into the interaction loop. Project tutors trained in both QtA and in the tutorial subject matter served as the wizards. During a session tutors were responsible for accepting, overriding, and/or authoring system actions. Tutor wizards were also responsible for setting the current dialogue frame to indicate which of the learning goals was currently in focus. Students talked to MyST via microphone while MyST communicates using Text-to-Speech (TTS) in the WoZ setting. A typical MyST session revolves around a single FOSS lesson and lasts approximately 15 minutes. To obtain a dialogue transcript, tutor moves are taken directly from the system logfile, while student speech is manually transcribed from audio. In addition to the dialogue text, MyST records additional information such as timestamps and the current dialogue frame (i.e. learning goal). In total we make use of transcripts from 122 WoZ dialogues covering 10 units on magnetism and electricity and 2 in measurement and standards.

3.2 Dialogue Annotation

Lesson-independent analysis of dialogue requires a level of abstraction that reduces a dialogue to its underlying actions and intentions. To address this need we use the Dialogue Schema Unifying Speech and Semantics (DISCUSS) (Becker et al.,

[1] http://www.fossweb.com

2011), a multidimensional dialogue move taxonomy that captures both the pragmatic and semantic interpretation of an utterance. Instead of using one label, a DISCUSS move is a tuple composed of three dimensions: *Dialogue Act*, *Rhetorical Form*, *Predicate Type*. Together these labels account for the action, function, and content of an utterance. This scheme draws from past work in task-oriented dialogue acts (Bunt, 2009; Core and Allen, 1997), tutorial act taxonomies (Pilkington, 1999; Tsovaltzi and Karagjosova, 2004; Buckley and Wolska, 2008; Boyer et al., 2009b) discourse relations (Mann and Thompson, 1986) and question taxonomies (Graesser and Person, 1994; Nielsen et al., 2008).

Dialogue Act (22 tags): The dialogue act dimension is the top-level dimension in DISCUSS, and its values govern the possible values for the other dimensions. Though the DISCUSS dialogue act layer seeks to replicate the learnings from other well-established taxonomies like DIT++ (Bunt, 2009) or DAMSL (Core and Allen, 1997) wherever possible, the QtA style of pedagogy driving our tutoring sessions dictated the addition of two tutorial specific acts: marking and revoicing. A *mark* act highlights key words from the student's speech to draw attention to a particular term or concept. Like with marking, *revoicing* keys in on student language, but instead of highlighting specific words, a *revoice* act will summarize or refine the student's language to bring clarity to a concept.

Rhetorical Form (22 tags): Although the dialogue act is useful for identifying the speaker's intent, it gives no indication of how the speaker is advancing the conversation. The rhetorical form refines the dialogue act by providing a link to its surface form realization. Consider the questions "What is the battery doing?" and "Which one is the battery?". They would both be labeled with *Ask* dialogue acts, but they elicit two very different kinds of responses. The former, which elicits some form of description, would be labeled with a *Describe* rhetorical form, while the latter is seeking to *Identify* an object. Similarly an *Assert* act from a tutor could be coupled with a *Describe* rhetorical form to introduce new information or with a *Recap* to reconvey a major point.

Predicate Type (19 tags): Beyond knowing the propositional content of an utterance, it is useful to know how the entities and predicates in a response relate to one another. A student may mention several keywords that are semantically similar to the learning goals, but it is important for a tutor to recognize whether the student's language provides a deeper description of some phenomena or if it is simply a superficial observation. The Predicate Type aims to categorize the semantic relationships a student may talk about; whether it is a *Procedure*, a *Function*, a *Causal Relation*, or some other predicate type.

Reliability Metric	DA	RF	PT
Cohen's Kappa	0.75	0.72	0.63
Exact Agreement	0.80	0.66	0.56
Partial Agreement	0.89	0.77	0.68

Table 1: Inter-annotator agreement for DISCUSS types (DA=Dialogue Act, RF=Rhetorical Form, PT=Predicate Type)

3.2.1 Annotation

All transcripts used in this experiment have been annotated with DISCUSS labels at the turn level. A reliability study using 15% of the transcripts was conducted to assess inter-rater agreement of DISCUSS tagging. This consisted of 18 doubly annotated transcripts comprised of 828 dialogue utterances.

To assess inter-rater reliability we use Cohen's Kappa (κ) (Carletta, 1996). Because DISCUSS permits multiple labels per instance, we compute a κ value for each label and provide a mean for each DISCUSS dimension. To get an additional sense of agreement, we use two other metrics: exact agreement and partial agreement. For each of these metrics, we treat each annotators' annotations as a per class bag-of-labels. For exact agreement, each annotators' 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 1 lists all three metrics broken down by DISCUSS dimension. The κ 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 diffi-

culty in labeling each dimension, and the agreement as a whole illustrates the open-endedness of the task.

3.3 Question Authoring

While the long-term plan for this work is to integrate fully automatic question generation into a tutoring system, for this study we opted to use manually authored questions. This allows us to remain focused on learning to identify context appropriate questions rather than confounding our experiments with issues of question grammaticality and well-formedness. Even though using multiple authors would provide greater diversity of questions, to avoid repeated effort and to maintain consistency in authoring we trained a single question author in both the FOSS material and MyST QtA techniques. Although he was free to author any question he found appropriate, our 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, we taught him to consider how QtA moves such as *Revoicing*, *Marking*, or *Recapping* could alter otherwise similar questions. To minimize the risk of rater bias, we explicitly told our author to avoid using positive feedback expressions such as "Good job!" or "Great!". Table 2 illustrates how the combinations of DISCUSS labels, QtA tactics, and dialogue context drives 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. 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 allowed us to focus on follow-up style questions. Because these question authoring contexts came from actual tutorial dialogues, we also extracted the original turn provided by the tutor, and we filtered out turns that did not contain questions related to the lesson content. Our corpus has 205 question authoring contexts comprised of 1025 manually authored questions and 131 questions extracted from the original transcript yielding 1156 questions in total.

3.4 Ratings Collection

To rate questions, we enlisted the help of four tutors who had previously served as project tutors and wizards. The 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.

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 we instead ask raters to simultaneously score all candidate questions. Because we did not want to bias raters, we did not specify specific criteria for question quality. Instead we instructed the raters 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 we collected ratings for 1156 question representing a total of 205 question contexts distributed across 30 transcripts.

3.4.1 Rater Agreement

Because these judgments are subjective, a key challenge in this work centers on understanding to what degree the tutors agree with one another. Since our goal is to rank questions and not to score questions, we convert each tutors scores for a given context into a rank-ordered list. To compute inter-rater agreement in ranking, we use Kendall's-Tau (τ) rank correlation coefficient. 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 typ-

...

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.*

	Candidate Question	Frame	Element	DISCUSS
Q1	Roll over the switch and then in your own words, tell me again what a complete or closed circuit is all about.	Same	Same	Direct/Task/Visual Ask/Describe/Configuration
Q2	How is this circuit setup? Is it open or closed?	Same	Same	Ask/Select/Configuration
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?	Diff	Diff	Assert/Recap/Proposition Direct/Task/Visual Ask/Describe/Function
Q4	You said something about the motor spinning in a complete circuit. Tell me more about that.	Same	Same	Revoice/None/None Ask/Elaborate/CausalRelation

Table 2: Example dialogue context snippet and a collection of candidate questions. The frame, element, and DISCUSS columns show how the questions vary from one another.

ically lower when the range of possible choices is narrow as it is in this task. To get a single score we average τ values across all sets of questions (contexts) and all pairs of raters. The mean value for all pairs of raters and contexts is $\tau = 0.1478$. The inter-rater statistics are shown in table 3. While inter-rater agreement is fairly modest, we do see lots of variation between different pairs of tutors. Additionally, we found that a pair of raters agreed on the top rated question 33% of the time. This suggests that despite their common training and experience, the raters may be using different criteria in rating.

To assess the tutors' internal consistency, we had each tutor re-rate 60 sets of questions approximately two months after their first trial, and we computed self-agreement Kendall's-τ values using the method above. These statistics are listed in the bottom row of table 3. In contrast with the inter-rater agreement, self-agreement is much more consistent giving further evidence for a difference in criteria. Together self and inter-rater agreement help bound expected system performance in ranking.

	rater A	rater B	rater C	rater D
rater A	X	0.2590	0.1418	0.0075
rater B	0.2590	X	0.1217	0.2370
rater C	0.1418	0.1217	X	0.0540
rater D	0.0075	0.2370	0.0540	X
mean	0.1361	0.2059	0.1058	0.0995
self	0.4802	0.4022	0.2327	0.3531

Table 3: Inter-rater rank agreement (Kendall's-τ). The bottom row is the self-agreement for contexts they rated in two separate trials.

4 Automatic Ranking

Because we are more interested in learning to predict which questions are more suitable for a given tutoring scenario than we are in assigning specific scores to questions, we approach the task of question selection as a ranking task. To create a gold-standard for training and evaluation we first need to convert the collective ratings for a set of questions 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 we create a single rank-order by tabulating pairwise wins for all pairs of questions $q_i, q_j, (i \neq j)$ within a given dialogue context C. If $rating(q_i) > rating(q_j)$, questions q_i receives a win. This is summed across all raters for the 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 we then train a pairwise classifier to learn a preferences function (Cohen et al., 1998) that determines if one question has a better rank than another. For each question q_i within a context C, we construct a vector of features ϕ_i. For a pair of questions q_i and q_j, we then create a new vector using the difference of features: $\Phi(q_i, q_j, C) = \phi_i - \phi_j$. For training, if $rank(q_i) < rank(q_j)$, the classification is positive otherwise it is negative. To account for the possibility of ties, and to make the difference measure appear symmetric, we train both combinations (q_i, q_j) and (q_j, q_i). During decoding, we run the trained classifier on all pairs and tabulate wins using the approach described above.

For our experiments we train pairwise classifiers using Mallet's Maximum Entropy (McCallum, 2002) and SVM^{Light}'s Support Vector Machines models (Joachims, 1999). We also use SVM^{Rank} (Joachims, 1999), which performs the same maximum margin separation as SVM^{Light}, but uses Kendall's-τ as a loss function to optimize for rank ordering. We run SVM^{Rank} with a linear kernel and model parameters of $c = 2.0$ and $\epsilon = 0.0156$. For MaxEnt, we use Mallet's default model parameters. 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, we train and evaluate models by incrementally adding additional information extracted from the DISCUSS annotation.

4.1 Features

When designing features for this task, we wanted 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, lesson relevance, contextual relevance. The subsections below detail the motivations and intuitions behind these factors.

4.1.1 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 we use the *bag-of-part-of-speech-tags (POS)* features to provide another aspect of the question's structure.

4.1.2 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, we hypothesize 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 we extract features for unigram and bigram overlap of words, word-lemmas, and part-of-speech tags between the pairs 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 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.

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 4: Example learning goals

4.1.3 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 allows us to bridge between the question's semantics and pragmatically and focus 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. We capture the question's dialogue cohesiveness with binary features indicating whether or not the question's RF and PT match those found in the previous student and tutor turns.

4.1.4 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.

We model the conditional relevance of moves by computing dialogue act transition probabilities from our corpus of DISCUSS annotated tutorial dialogues. Although DISCUSS allows multiple tags per dialogue turn, we simplify probability calculations 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 we use different subsets of the tuple to compute the transition probabilities listed in equations 1-3. All probabilities are computed using Laplace-smoothing. When extracting features, we sum the log of the probabilities 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. To account for student progress in question asking, we compute the conditional probability of a DISCUSS label given the percentage of elements filled in the current dialogue frame (equation 4). This progress percentage is discretized into bins of 0-25%, 25-50%, 50-75%, and 75-100%.

$$p(DA, RF, PT_{question}|DA, RF, PT_{stud.\ turn}) \quad (1)$$

$$p(DA, RF_{question}|DA, RF_{student\ turn}) \quad (2)$$

$$p(PT_{question}|PT_{student\ turn}) \quad (3)$$

$$p(DA, RF, PT_{ques.}|\%\ elements\ filled) \quad (4)$$

4.2 Evaluation

To evaluate our systems' performance in ranking, we use two measures commonly used in information retrieval: the Mean Kendall's-τ measure described in section 3.4.1 and Mean Reciprocal Rank (MRR). MRR is the average of the multiplicative inverse of the rank of the highest ranking question across all contexts. To account for ties we use the Tau-b variant of Kendall's-τ , and for MRR we compute reciprocal rank by averaging the system rankings for all of the questions tied for first. To obtain a gold-standard ranking for comparison, we combine individual raters' ratings using the approached described in section 4.

5 Results and Discussion

We trained several models to investigate how different feature classes influence overall performance in ranking. The results for these experiments are listed in Table 5. Because we found comparable performance between MaxEnt and SVM^{Light}, we only report results for MaxEnt and SVM^{Rank} models. In addition to MRR and Kendall's-τ , we list the number of concordances and discordances in pairwise classification to give the reader another sense of the accuracy associated with rank agreement.

Random Baseline: On average, assigning random ranks will yield mean τ=0 and MRR=0.408.

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

Table 5: System scores by feature set and 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 4.1.1 and 4.1.2. **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 4.1.3. **MATCH** refers specifically to the RF and PT match features. **CONTEXT** refers to the Contextualized DISCUSS features described in section 4.1.4. The best scores for each column appear in boldface.

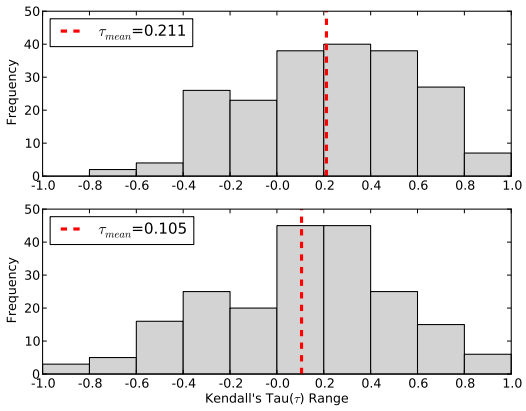

Figure 1: Distribution of per-context Kendall's-τ values for the top-scoring system (top), and the baseline system (bottom).

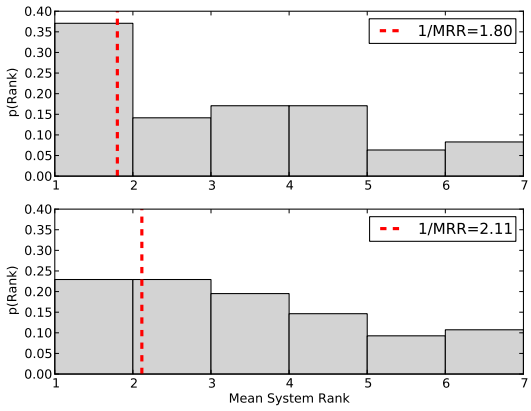

Figure 2: Distribution of per-context system ranks for the highest rated question for the top-scoring system (top), and the baseline system (bottom). These ranks are the inverse of the reciprocal rank used to calculate MRR.

Baseline System: Our 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 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 our ranking models to their highest level of agreement with $\tau = 0.211$ using MaxEnt and MRR=0.555 using SVM^{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 44/205 (21.4%) contexts. While the system with the best MRR score, correctly chooses the top-ranked question in 71/205 (34.6%) contexts – 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 1 and 2 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, we evaluated our human raters on 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 we remove ratings 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 our best-performing system.

Looking at the impact of learning algorithms we see that SVM^{Rank} tends to perform better on MRR while the pairwise maximum entropy models yield higher τ's. 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.

6 Conclusions and Future Work

We have 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. Supplementing our baseline surface form and lexical similarity features with additional features extracted from the dialogue context and DISCUSS dialogue act annotation improves system performance in ranking to a level on par with expert 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.

In the near future we plan to train models on individual tutors to investigate which factors drive individual preferences in question asking. We also plan to characterize system performance using automatically labeled DISCUSS annotation. Lastly, we feel these results provide a natural starting point to explore automatic generation of questions from the DISCUSS dialogue move representation.

Acknowledgments

This work was supported by grants from the NSF (DRL-0733322, DRL-0733323), the IES (R3053070434) and the DARPA GALE program (Contract No. 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.

References

Manish Agarwal and Prashanth Mannem. 2011. Automatic gap-fill question generation from text books automatic gap-fill question generation from text books automatic gap-fill questions from text books. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*.

I. L. Beck, M. G. McKeown, J. Worthy, C. A. Sandora, and L. Kucan. 1996. Questioning the author: A year-long classroom implementation to engage students with text. *The Elementary School Journal*, 96(4):387–416.

L. Becker, R. D. Nielsen, and W. Ward. 2009. What a pilot study says about running a question generation challenge. In *Proceedings of the Second Workshop on Question Generation*, Brighton, England, July.

L. Becker, W. Ward, S. van Vuuren, and M. Palmer. 2011. Discuss: A dialogue move taxonomy layered over semantic representations. In *In Proceedings of the International Conference on Computational Semantics (IWCS) 2011*, Oxford, England, January 12-14.

K.E. Boyer, E.Y. Ha, M. Wallis, R. Phillips, M.A. Vouk, and J.C. Lester. 2009a. Discovering tutorial dialogue

strategies with hidden markov models. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED '09)*, pages 141–148, Brighton, U.K.

K.E. Boyer, W.J. Lahti, R. Phillips, M. D. Wallis, M. A. Vouk, and J. C. Lester. 2009b. An empirically derived question taxonomy for task-oriented tutorial dialogue. In *Proceedings of the Second Workshop on Question Generation*, pages 9–16, Brighton, U.K.

M. Buckley and M. Wolska. 2008. A classification of dialogue actions in tutorial dialogue. In *Proceedings of COLING 2008*, pages 73–80. ACL.

H. C. Bunt. 2009. The DIT++ taxonomy for functional dialogue markup. In *Proc. EDAML 2009*.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):pp. 249–254.

M. Chi, P. Jordan, K. VanLehn, and M. Hall. 2008. Reinforcement learning-based feature selection for developing pedagogically effective tutorial dialogue tactics. In Ryan S. Baker, Tiffany Barnes, and Joseph Becker, editors, *Proceedings of the 1st International Conference on Educational Data Mining*, pages pp258–265.

M. Chi, P. W. Jordan, K. VanLehn, and D. J. Litman. 2009. To elicit or to tell: Does it matter? In *Artificial Intelligence in Education*, pages 197–204.

M. Chi, K. VanLehn, and D. Litman. 2010. Do micro-level tutorial decisions matter: Applying reinforcement learning to induce do micro-level tutorial decisions matter. In Vincent Aleven, Judy Kay, and Jack Mostow, editors, *Preceedings of the 10th Internation Confernce on Intelligent Tutoring Systems (ITS 2010)*.

William W. Cohen, Robert E. Schapire, and Yoram Singer. 1998. Learning to order things. In *Advances in Neural Information Processing Systems 10 (NIPS 1998)*.

A. Collins and A. Stevens. 1982. Goals and methods for inquiry teachers. *Advances in Instructional Psychology*, 2.

M. G. Core and J.F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium*, pages 28–35.

A.C. Graesser and N.K. Person. 1994. Question asking during tutoring. *American Educational Research Journal*, 31:104–137.

A.C. Graesser, P. Chipman, B.C Haynes, and A. Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education*, 48:612–618.

M. Heilman and N. A. Smith. 2010. Good question! statistical ranking for question generation. In *Proceedings of NAACL/HLT 2010*.

T. Joachims. 1999. Making large-scale svm learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

J.H. Kim, M. Glass, R. Freedman, and M.W. Evens. 2000. Learning the use of discourse markers in tutorial dialogue learning the use of discourse markers in tutorial dialogue. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*.

D. Litman and S. Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *Companion Proceedings of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*.

W.C. Mann and S.A Thompson. 1986. Rhetorical structure theory: Description and construction of text structures. In *In Proceedings of the Third International Workshop on Text Generation*, August.

A. K. McCallum, 2002. *MALLET: A Machine Learning for Language Toolkit.* http://mallet.cs.umass.edu.

R. D. Nielsen, J. Buckingham, G. Knoll, B. Marsh, and L. Palen. 2008. A taxonomy of questions for question generation. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, September.

R.M. Pilkington. 1999. Analysing educational discourse: The discount scheme. Technical Report 99/2, Computer Based Learning Unit, University of Leeds.

Owen Rambow, Monica Rogati, and Marilyn A. Walker. 2001. Evaluating a trainable sentence planner for a spoken dialogue system evaluating a trainable sentence planner for a spoken dialogue system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*.

C.P. Rose, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein. 2001. A comparative evaluation of socratic versus didactic tutoring. In *Proceedings of Cognitive Sciences Society*.

C.P. Rose, D. Bhembe, S. Siler, R. Srivastava, and K. VanLehn. 2003. The role of why questions in effective human tutoring. In *Proceedings of Artificial Intelligence in Education (AIED 2003)*.

D. Tsovaltzi and E. Karagjosova. 2004. A view on dialogue move taxonomies for tutorial dialogues. In *Proceedings of SIGDIAL 2004*, pages 35–38. ACL.

D. Tsovaltzi and C. Matheson. 2001. Formalising hinting in tutorial dialogues. In *In EDILOG: 6th workshop on the semantics and pragmatics of dialogue*, pages 185–192.

K. VanLehn, A.C. Graesser, G.T. Jackson, P. Jordan, A. Olney, and C.P. Rose. 2007. When are tutorial dialogues more effective than reading? *Cognitive Science*, 31(1):3–62.

Marilyn A. Walker, Owen Rambow, and Monica Rogati. 2001. SPOT: A trainable sentence planner. In *Proceedings of the North American Meeting of the Association for Computational Linguistics (NAACL)*.

A. Ward, D. Litman, and M. Eskenazi. 2011a. Predicting change in student motivation by measuring cohesion between predicting change in student motivation by measuring cohesion between tutor and student. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 136–141.

W. Ward, R. Cole, D. Bolaños, C. Buchenroth-Martin, E. Svirsky, S. van Vuuren, T. Weston, J. Zheng, and L. Becker. 2011b. My science tutor: A conversational multi-media virtual tutor for elementary school science. *ACM Transactions on Speech and Language Processing (TSLP)*, 7(4), August.

W. Ward. 1994. Extracting information from spontaneous speech. In *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*.

Xuchen Yao. 2010. Question generation with minimal recursion semantics. Master's thesis, Saarland University.