Evaluating Questions in Context

Lee Becker and **Martha S. Palmer** and **Sarel van Vuuren**
The Center for Computational Language and Education Research
University of Colorado Boulder
{lee.becker, sarel.vanvuuren, martha.palmer}@colorado.edu

Wayne H. Ward
Boulder Language Technologies
wward@bltek.com

Abstract

We present an evaluation methodology and a system for ranking questions within the context of a multimodal tutorial dialogue. Such a framework has applications for automatic question selection and generation in intelligent tutoring systems. To create this ranking system we manually author candidate questions for specific points in a dialogue and have raters assign scores to these questions. To explore the role of question type in scoring, we annotate dialogue turns with labels from the DISCUSS dialogue move taxonomy. Questions are ranked using a SVM-regression model trained with features extracted from the dialogue context, the candidate question, and the human ratings. Evaluation shows that our system's rankings correlate with human judgments in question ranking.

1 Introduction

There is currently a tide of interest in developing applications that make use of question generation systems. While much of the effort has largely focused on generation of questions from text, most applications of question generation can be considered dialogue systems where the question is generated based on not only the source material, but also on the series of interactions leading to the current question generation context. Nielsen (2008) describes question generation as a multistage process consisting of three sub-tasks: Target Concept Identification, Question Type Determination, and Question Realization, which translates into determining what to talk about, determining how to talk about it, and lastly creating a question from these representations. Furthermore Vanderwende (2008) states the process of identifying which question to ask is as critical as generating the question itself.

This paper focuses on question selection, a task closely related to question type determination, and is motivated by the overarching goal of learning strategies for scaffolding and sequencing moves within an intelligent tutoring system. We treat question selection as a task of scoring and ranking candidate questions. Rather than optimizing and scoring for grammaticality or correctness of the question's surface form realization, we are evaluating with respect to a question's utility and appropriateness for a specific point in a tutorial dialogue. Because these decisions are highly dependent on the curriculum and choice of pedagogy, we ground our investigation into question ranking within a single paradigm. Specifically we are learning to rank questions and interactions stylistically similar to those found within the My Science Tutor (MyST) intelligent tutoring system (Ward et al. 2011).

MyST is 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. MyST's curriculum is based on the Full Option Science System (FOSS) [1] a proven research-based science curriculum that has been widely deployed in American schools for over a decade. FOSS consists of sixteen teaching and learning modules covering life science, physical science, earth and space science, scientific reasoning, and technology. For this study, we limit our coverage of FOSS to investigations about magnetism and electricity.

In the following sections we describe related work in question ranking and dialogue move selection, detail the process of data collection, describe our approach to question ranking, share our experimental results, and close with conclusions and suggestions for future work.

2 Connections to Prior Work

While there is a growing body of research in question generation for educational texts, few focus on scoring and ranking of questions in the context of a tutorial dialogue. Heilman and Smith (2010b; 2010a) use a statistical model to rate the quality of questions; however, this system is tuned more for rating grammaticality and syntactic correctness

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

than for tutoring sequencing and scaffolding. Existing research in dialogue management borders on this kind of tutorial decision making. Chi et al. (2009) uses reinforcement learning to optimize between two specific questioning tactics – eliciting and telling; however exploring the full array of tutoring options can quickly become intractable. Overgeneration and ranking for dialogue move generation presents an opportunity to bridge between the lower-level generation details and higher level dialogue planning. Varges (2006) utilizes features of the dialogue to influence generation and then applies heuristics to rank of dialogue moves in a restaurant domain. Because our goal is to learn effective tutoring strategies, we instead instead aim to learn ranking models directly from human judgement data.

This paper's contributions are as follows:

- We develop a methodology for evaluating Question Generation in the context of a dialogue.
- We apply statistical machine learning to create a model for scoring and ranking candidate questions.
- We lay the foundation for future investigation into the utility of a rich, multi-level dialogue representation for scoring candidate questions.

3 Data Collection

MyST Logfiles and Transcripts

To gather natural interactions during MyST's development stage, we ran Wizard-of-Oz (WoZ) experiments that inserted a human tutor into the interaction loop. Project tutors trained in both QtA and in the tutorial subject matter served as the wizards. During a session they were responsible for accepting, overriding, and/or authoring system actions. They were also responsible for managing which of the learning goals was currently in focus. Over the past year, we have also been collecting assessment data using MyST's standalone condition. In both conditions students talk to MyST via microphone, while MyST communicates using Text-to-Speech (TTS) in the WoZ setting and pre-recorded audio in the standalone setting. A typical MyST session covers on part of a FOSS investigation and lasts approximately 15 minutes. For this work, we exclusively use data collected from the WoZ experiments.

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 logs additional information such as timestamps, the current dialogue frame (target concept) and frame-element (subpart of a target concept). The transcripts used in this study represents a small subset of all the data we have collected and consists of 103 WoZ dialogues covering 9 different lessons on magnetism and electricity.

Dialogue Annotation

DISCUSS 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. 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 four dimensions: *Dialogue Act*, *Rhetorical Form*, *Predicate Type* and *Semantic Roles*. Together these labels provide a full account of the action and meaning 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. 2009) discourse relations (Mann and Thompson 1986) and question taxonomies (Graesser and Person 1994; Nielsen et al. 2008).

Dialogue Act 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 *marking*, *revoicing* also 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 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 cue 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, eliciting some form of description would be labeled with an *Describe* rhetorical form, while the latter is seeking to *Identify* an object.

Predicate Type 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 state 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 type.

Semantic Roles The MyST system models a lesson's key concepts as propositions which are realized as semantic frames. For MyST natural language understanding, these frames serve as the top-level nodes for a manually written semantic grammar used by the Phoenix parser (Ward 1994). While the MyST system uses Phoenix semantic frames to drive dialogue behavior, generalizing behavior across lessons requires a more domain-independent semantic representation. DISCUSS adapted and expanded VerbNet's (Kipper, Dang, and Palmer 2000; Schuler 2005) set of

Reliability Metric	DA	RF	PT
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)

thematic roles to the needs of our tutoring domain. The shallow semantics provides portability while maintaining descriptiveness.

Annotation All transcripts used in this experiment have been annotated with DISCUSS labels at the turn level. A reliability study over 16.5% of the transcripts was conducted to assess inter-rater agreement of DISCUSS tagging. This consisted of 18 doubly annotated transcripts totaling to 828 dialogue utterances.

Because DISCUSS permits multiple labels per instance, we present two metrics for reliability: 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. Agreement credit in partial agreement is defined as the number of intersecting labels divided by the total number of unique labels. Agreement numbers are shown in table 1. While the dialogue act and rhetorical form agreement are relatively high, the low agreement in the predicate type labeling reflects the difficulty and open-endedness of the task.

Question Authoring

While it would be possible to create a system that generates candidate questions, we opted to use manual authoring of questions to avoid conflating issues of grammaticality with question appropriateness. To collect questions we trained a linguist to author questions. The question author was presented with much of the same information available to the MyST system including the entire dialogue history up to the decision point and the current frame in focus. Instead of leaving the task purely open-ended, we asked our authors to generate questions by considering permutations in DISCUSS representation, the target frame, and frame-element. The author was instructed to consider how QtA acts such as *Revoicing*, *Marking*, or *Recapping* could alter otherwise similar questions. Other authoring decisions included choosing between whether to wrap-up or to remain in the current topic. Table 2 illustrates how an author may explore the combinations of DISCUSS labels, QtA tactics, and topic choices to produce candidate questions for a given context.

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 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.

Ratings Collection

To rate the questions, we utilized workers from Amazon's Mechanical Turk[2] crowdsourcing service. As with question authoring, the workers were presented with tutorial dialogue history preceding the question decision point, and a list of 6 candidate questions (5 manually authored, 1 taken from the original transcript). To give additional context, raters were also presented a list of the lessons' learning goals and were given links to view the interactive visuals displayed in the tutoring system.

Previously Becker, Nielsen, and Ward (2009) found poor agreement when rating individual questions in isolation. To decrease the task's difficulty, we instead ask raters to simultaneously score all candidate questions. We also instructed the raters to consider factors such as "whether or not it is better to move on or to remain with the current line of questioning, whether the question seems out of place, or whether it assists the student's understanding of the learning goals." Scores are collected using an ordinal 10-point scale ranging from 1 (lowest/worst) to 10 (highest/best). Rating collection is still a work in progress. At the time of this writing, we have collected ratings for 288 of the 1156 questions, representing a total of 51 question contexts across 8 transcripts.

4 Automatic Ranking

Our ranking model is built using SVM regression from SVM^{light}'s (Joachims 1999) to rank candidate questions similar to the approach used by Heilman and Smith (2010b). The regression is trained on the average score of the human raters, using the default SVM parameters. Features are extracted using the ClearTK (Ogren, Wetzler, and Bethard 2008) statistical NLP framework. Training and evaluation is done using 8-fold cross validation partitioned to ensure questions from the same transcript are not in both the training and validation set.

The following subsections list the motivations and descriptions of many of the features we expect to be important for question ranking and selection. Due to time constraints, the system evaluated in this paper does not make use of the Dialogue Context and DISCUSS features, but we hope to present additional findings at the workshop.

Basic Features

Questions that are too wordy or too terse may score poorly with raters. Additionally, the verbosity of student answers may reflect how raters score the follow-up questions. To capture this, we extract length features from the student's response and the candidate question.

[2] https://www.mturk.com/

. . .

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.

Lexical and Syntactic Features

While it is common to include a bag-of-words feature in many NLP tasks including document classification and sentiment analysis, this approach is less useful when dealing with both a small corpus and multiple domains. To circumvent potential issues of sparsity, we approximate word features with part-of-speech (POS) tag features including a question's POS-tag frequency and distributions. To drive towards question type, we also have features indicating the presence of Wh-words (who, what, why, where, when, how, which, etc. . .).

Semantic Similarity Features

In QtA, tutoring actions are guided by the goal of eliciting student responses that address the learning goals for the lesson. Additionally, common QtA moves involve highlighting or paraphrasing of student speech. To detect how responses influence a question's score, we use several semantic similarity measures between:

- The student's response and the candidate question
- The student's response and the preceding tutor question
- The student's response and the text of the learning goals

For these experiments we use unigram and bigram overlap of words, word-lemmas, and part-of-speech tags as a first pass at measuring semantic similarity.

Dialogue Context Features

This set includes features such as the number of turns spent in the current frame, the number of turns spent on the current frame-element, the cumulative fraction of elements filling in the current frame, the fraction of elements filling the current frame by the last turn.

DISCUSS Features

This set of features makes use of the DISCUSS representation for both the candidate question and for the student and tutor dialogue moves. Basic DISCUSS features include the dialogue acts, rhetorical forms, and predicate types contained within 1) the tutor's initiating question and 2) the student's response.

5 Evaluation and Results

Evaluation

Our long term goal in designing a suitable evaluation for this task is to enable evaluation of fully automatic question generation systems. In the closer term, the experimental design is focused on creating a framework that will allow analysis of what features and representations are best suited for this task.

Ranking questions in context is a highly subjective task, and with only 2-3 raters per context, there is no single ground truth ranking for evaluation. In a previous study on scoring questions (Becker, Nielsen, and Ward 2009), we calculated correlation between the score produced by the system and the mean score for a candidate question using Pearson's correlation coefficient; however, this only reflects per question similarity of scores and does not address their relative rankings. A similar limitation holds for direct evaluation using metrics such as the regression function's mean-squared-error.

Because we are more interested in the system's overall ability to rank questions against one another, we opted to frame evaluation in terms of rater agreement. Rather than evaluating the system on gold-standard data, we treat the system as one of several raters, and compare system-human agreement to human-human agreement. To do this we make use of two statistics Kendall's tau (τ) coefficient (Kendall 1938) and the Mann-Whitney U test (Mann and Whitney 1947). Kendall's tau is a statistic to measure the correlation between two pairs of rankings, and the Mann-Whitney U test is a non-parametric significance test. We use these statistics to gauge whether system-human ranking agreement is equal or better than human-human rankings. We also generated random rankings to ensure that human-human ranking agreement was statistically better than from human-random ranking agreement. The evaluation procedure is as follows:

Algorithm 1 Evaluation Procedure

Individual τ statistics are accumulated to collect distributions of agreement. The final U-test indicates the significance in overlap between the distributions.

```
for all Question Contexts do
    for all (r1, r2) in Human-Human Ranking Pairs do
        append KendallTau(r1,r2) to agreeHH
    end for
    for all (r1, r2) in System-Human Ranking Pairs do
        append KendallTau(r1,r2) to agreeSH
    end for
end for
U := MannWhitney(agreeHH, agreeSH)
```

Results and Discussion

Ranking agreement statistics are shown in table 3 and their distributions are illustrated in figure 1. Results of the Mann-Whitney U test are shown in table 4. The distributions of Human-Human ranking agreements and System-Human ranking agreements showed no statistical difference ($p > 0.05$), and the distributions for Human-Human and Random-Human agreement differed significantly ($p < 0.05$)

Agreement (τ)	n	Mean	Std	Median
Human-Human	226	0.084	0.388	0.138
System-Human	128	0.115	0.375	0.138
Random-Human	128	0.017	0.329	0.036

Table 3: Agreement (Kendall's Tau) Statistics: τ=1 indicates perfect agreement, τ=-1 perfect disagreement, and τ=0 no agreement

Agreement Dists.	n_1	n_2	U	p
System-Human	128	226	25538	0.357
Random-Human	128	226	12054	0.004

Table 4: Mann-Whitney U test statistics: p can take values between 0 and 1. A p of 0.5 indicates exact overlap between distributions. Values close to 0 and 1 indicate complete separation of distributions.

The results above suggest that untrained human raters have slight agreement in ranking questions, but on average do better than random selection. We also find it encouraging that without using the more complex dialogue context and DISCUSS features, we were able to create a system that can replicate the ratings of novice raters. However, we do not believe this finding eliminates the need for such features. Instead, it may suggest that novice raters cue in on more superficial features when scoring questions and may not take into account more complex behaviors like sequencing and question types. Future ablation studies comparing experts and novices should help to identify the importance of more complex features.

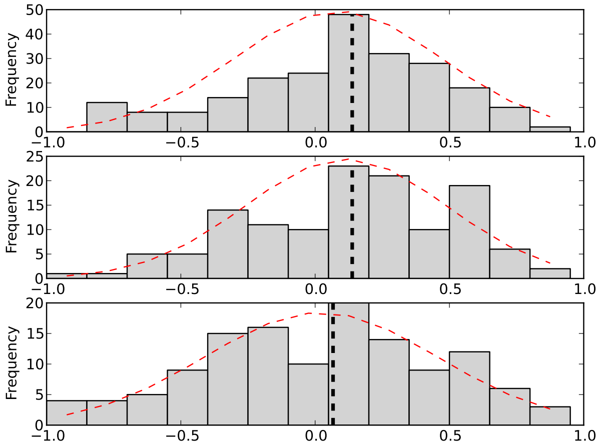

Figure 1: Distribution of Ranking Agreements (Kendall's Tau scores): Human-Human Agreement (top), System-Human Agreement (middle), Random-Human Agreement (bottom)

When analyzing inter-rater agreement, one should also consider the raters' inexperience with the QtA style of teaching. The following comments below highlight some raters' negative reactions to QtA specific prompts:

> *I think with children, you should use more simple questions that doesn't make them try and reason with what they are saying since they can't grasp that concept yet.*
>
> *The phrase "What's up with that" in Q5 *really* grates on me.*
>
> *Questions seemed more successful that had concrete answers and less when they were observational and open end.* (sic)

These comments illustrate the challenge in dealing with pedagogical bias when running a tutoring oriented question generation evaluation, and they provide anecdotal evidence that using raters versed in QtA and the FOSS curriculum should yield improved reliability in rating and increased performance in our models.

6 Conclusion and Future Work

We have introduced a new methodology for assessing the quality of question generation in the context of a tutorial dialogue, and have developed a system that automatically scores and ranks candidate questions. Analysis of question ratings shows that novice raters show slight agreement when ranking questions. The performance of our question ranking system demonstrates the feasibility of this task and shows that it is possible for system to produce rankings that correlate with human judgments.

In the near term, we will continue to refine this system by collecting additional ratings from novice raters (Mechanical Turk) and from expert tutors trained in QtA techniques and the FOSS curriculum. There is still much room for further improvement in system performance as the model does not account for any of the complexity associated with the dialogue. To address these needs we plant to annotate the candidate questions with DISCUSS annotation, to allow for extraction of dialogue-specific features. Additionally, we intend to investigate how more sophisticated measures of semantic similarity affect system performance.

The ability to rank questions is a key piece of functionality for automatic question generation, and it represents an important step toward our larger goal of automatically learning behaviors for intelligent tutoring systems from corpora and other resources. While, we applied our analyses to manually authored questions, this framework can be applied toward evaluation and ranking of automatically generated questions. For future work we plan to investigate how applying this approach in conjunction with DISCUSS question types can be combined with existing techniques in overgeneration and ranking of questions.

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

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

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

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

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

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

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

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

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

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

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

Heilman, M., and Smith, N. A. 2010b. Rating computer-generated questions with mechanical turk. In *In Proceedings of the NAACL/HLT workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*.

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

Kendall, M. 1938. A new measure of rank correlation. *Biometrika* 30(1-2):81–89.

Kipper, K.; Dang, H. T.; and Palmer, M. S. 2000. Class based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence AAAI-2000*.

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

Mann, H., and Whitney, D. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18(1):50–60.

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

Nielsen, R. D. 2008. Identifying key concepts for question generation. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Ogren, P. V.; Wetzler, P. G.; and Bethard, S. 2008. ClearTK: A UIMA toolkit for statistical natural language processing. In *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC)*.

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

Schuler, K. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. Dissertation, University of Pennsylvania.

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

Vanderwende, L. 2008. The importance of being important. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Varges, S. 2006. Overgeneration and ranking for spoken dialogue systems. In *Proceedings of the Fourth International Natural Language Generation Conference*, 17–19.

Ward, W.; Cole, R.; Bolanos, D.; Buchenroth-Martin, C.; Svirsky, E.; Van Vuuren, S.; Weston, T.; Zheng, J.; and Becker, L. 2011. My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM TSLP (in review)*.

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