Detecting Question-Bearing Turns in Spoken Tutorial Dialogues

Jackson Liscombe, Jennifer J. Venditti, Julia Hirschberg

Spoken Language Processing Group
Columbia University
New York City, NY, USA
{jaxin,jjv,julia}@cs.columbia.edu

Abstract

Current speech-enabled Intelligent Tutoring Systems do not model student question behavior the way human tutors do, despite evidence indicating the importance of doing so. Our study examined a corpus of spoken tutorial dialogues collected for development of ITSpoke, an Intelligent Tutoring Spoken Dialogue System. The authors extracted prosodic, lexical, syntactic, and student and task dependent information from student turns. Results of running 5-fold cross validation machine learning experiments using AdaBoosted C4.5 decision trees show prediction of student question-bearing turns at a rate of 79.7%. The most useful features were prosodic, especially the pitch slope of the last 200 milliseconds of the student turn. Student pre-test score was the most-used feature. Findings indicate that using turn-based units is acceptable for incorporating question detection capability into practical Intelligent Tutoring Systems.

Index Terms: Intelligent Tutoring Systems, prosody, question-asking behavior, machine learning.

1. Introduction

Well designed Intelligent Tutoring Systems (ITSs), educational software designed to tutor students using artificial intelligence, are known to increase student learning over classroom instruction alone. However, the learning gains achieved with current ITSs are still well below the gains observed with human tutors. One reason for this could be that current speech-enabled ITSs do not model student question behavior the way human tutors do. Research has shown that question-asking on the part of students is an important part of tutoring interaction; for example, [1] observed up to 30 student questions per hour. In current ITSs, though, the rate of questions initiated by students is much lower, most likely because the experience is still distinctly different from interaction with a human tutor. While some researchers have begun to explore ITSs that elicit more questions from students [2], we know of no ITS that attempts to identify student questions explicitly. Our ultimate goals are to monitor the behavior of student users of ITSs so as to support question-asking and to respond appropriately to such questions. To this end, we present results of experiments that automatically predict student turns containing questions, using features extracted from the student's speech in a corpus of one-on-one spoken tutorial dialogues. We briefly note further results of research into the prediction of the **function** of student questions.

2. Corpus

For this research, we examined a corpus of spoken tutorial dialogues collected by [3] at the University of Pittsburgh. This corpus was collected for the development of ITSpoke, an Intelligent Tutoring Spoken Dialogue System designed to teach principles of qualitative physics. While the ITSpoke corpus comprises 12 hours of recorded speech, for this study we use only 141 dialogues between one (male) tutor and 17 college students (7 female, 10 male), containing 5 hours of student speech. A typical dialogue consists of approximately 53 student turns, each averaging 2.5 seconds and 5 words in length. The total number of student turns in the corpus is approximately 7,500.

The recording procedure for each session was as follows. The student and tutor were seated in the same room but separated by a partition so that they could not see each other. They interacted via microphones and a graphical user interface. Each student was first asked to type an essay in response to a qualitative physics question. The tutor then read the essay and proceeded to tutor the student verbally until he determined that the student had successfully mastered the material; at which point, the student would retype the essay. The student and tutor were recorded with separate microphones and each channel was manually transcribed and segmented into turns. An excerpt of a dialogue from the corpus is shown in Figure 1.

	... 17.4 minutes into the dialogue ...
TUTOR:	What does the acceleration mean?
STUDENT:	That the object is moving through space?
TUTOR:	No. Acceleration means that object's velocity is changing.
STUDENT:	What?
TUTOR:	Object's velocity is changing.
STUDENT:	Uh-huh, and then once you release it the velocity remains constant.

Figure 1: A transcribed excerpt from the ITSpoke corpus of human-human spoken tutorial dialogues. Disfluencies have been eliminated and punctuation added for readability.

3. Annotation

For this study, the beginning and end of each question in the corpus were manually labeled. Each of the turns containing a question was further labeled as a *question-bearing turn*. In the work presented here, we are interested in determining whether a student turn as a whole contains a question or not, since ITSs typically interact with users in turn-based segments. In total, 1,030 questions were identified from 918 turns, a rate of roughly 25 per hour. This rate is consistent with other findings in one-on-one human tutoring, although it should be noted that the standard deviation is 13 questions per hour. Question behavior can be quite variable across students.

By adopting the turn as our unit of analysis we risk the masking of cues to questions by cues to other non-question phenomena present in the turn. However, we note that 70% of question-bearing turns consist entirely of the question itself. Of the remaining turns, 63% have questions that occur in turn-final position. In other words, 89% of question-bearing turns have questions that occur at the end the turn, indicating an area of the turn where questions are likely to occur.

4. Cues to Question-Bearing Turns

Many questions in Standard American English and other languages can be identified via lexical-syntactic cues; *e.g.,* [4], [5], [6]. For example, information-seeking questions often begin with one of the familiar *wh*-words (*e.g.,* 'what', 'who'). In addition, many questions exhibit inversion of the subject and auxiliary verb. These types of lexical-syntactic cues are clearly useful for question identification, though they do not identify all utterances that function pragmatically as questions. Pitch contour has long been considered important in this regard. In general, phrase-final rising intonation has been proposed for the identification of typical questions, specifically L* H-H% [7]. Such rising intonation may be most often present when a question otherwise would not differ from proper declarative statements, such as *yes-no* questions without inversion or declarative questions. Somewhat surprisingly, then, research has found that declarative questions are often intonationally equivalent to proper declaratives and that lexical-pragmatic cues are often necessary for differentiation [4]. As an example, utterances containing second person pronouns (*e.g.,* 'you') are more likely to be questions than those containing first person pronouns (*e.g.,* 'I') because, presumably, a speaker knows his or her own cognitive state but does not necessarily know that of the person he or she is speaking to. Other lexical-pragmatic cues suggested in the literature are utterance-initial particles (*e.g.,* 'oh').

Apart from lexical and intonational cues to questions, research also suggests that prosodic information other than pitch may play a role in question detection as well. For example, Shriberg et. al [8] found duration and pausing information to be more predictive than pitch in automatic question classification experiments. In fact, by automatically extracting prosodic features from utterances in the Switchboard corpus, they observed 74.21% accuracy in predicting questions versus non-questions. This was below the 83.65% accuracy using a language model trained on questions, though they observed increased performance (85.64%) when both sources of information were combined.

Motivated by the research presented above, we extracted several features from the speech signal in order to characterize student turns using prosodic, lexical, syntactic, as well as task and user-dependent information.

4.1. Prosodic Features

Most of the features we examined as potential indicators of question-bearing turns were prosodic features, including features associated with pitch, loudness, and rhythm. Acoustic processing was done in Praat, a program for speech analysis and synthesis [9]. Each prosodic feature was normalized by the speaker's mean value and recorded as a z-score.

We used fundamental frequency ($f0$) measurements to approximate overall **pitch** behavior. Features encapsulating pitch statistics – minimum, maximum, mean, and standard deviation – were calculated on all $f0$ information excluding the top and bottom 2% to eliminate outliers. Global pitch shape was approximated by calculating the slope of the all-points regression line over the entire turn. In addition, we wanted to isolate turn-final intonation shape. Accordingly, we smoothed and interpolated the $f0$ using built-in Praat algorithms and then isolated the last 200 milliseconds of the student turn over which we calculated the following $f0$ features: minimum, maximum, mean, standard deviation, slope of the line from first $f0$ point to the last, slope of all-points regression line, and the percent of rising slopes between each consecutive time points.

To examine the role of **loudness** we extracted the minimum, maximum, mean, and standard deviation of signal intensity, measured in decibels, over the entire student turn. In addition, we calculated the mean intensity over the last 200 milliseconds of each student turn, as well as the difference between the mean in the final region and the mean over the entire student turn.

Rhythmic features were designed to capture pausing and speaking rate behavior. We implemented a procedure to automatically identify pauses in student turns. The procedure isolates spans of silence 200 milliseconds or longer in length by using background noise estimation for each dialogue defined as the 75th quantile of intensity measurements over all non-student turns in that dialogue[1]. In the ITSpoke corpus we found there to be 1.62 pauses per student turn and the mean length of pauses to be 1.59 seconds. Pausing behavior in each student turn was represented as the number of pauses, the mean length of all pauses, the cumulative pause duration, and the percentage of time that pausing occupies relative to the entire student turn. Speaking rate was calculated by counting the number of voiced frames in the turn, normalized by the total number of frames in non-pause regions of the turn.

4.2. Non-prosodic Features

The remaining features we extracted from each student turn were non-prosodic. The **lexical** feature set comprises manually-transcribed word unigrams and bigrams uttered in each student turn. In addition to words with semantic content, we also included filled pauses, such as 'um' and 'uh'. To capture **syntactic** information we applied the Brill part-of-speech (POS) tagger, trained on the Switchboard corpus, to the lexical transcriptions of student turns. Syntactic features consist of POS unigrams and bigrams.

The remaining features were meant to capture knowledge about the student not present in either the aural or linguistic channels and are referred to as the **student and task dependent** feature set. Included in this set are: the score the student received on a physics test taken before the tutoring session (pre-test score), the gender of the student, the hand-labeled correctness of the student

[1] We refer the reader to [10] for a more detailed description of the algorithm.

Feature Set	Accuracy
none (majority class baseline)	50.0%
prosody: rhythmic	52.6%
student and task dependent	56.1%
prosody: loudness	61.8%
syntactic	65.3%
lexical	67.2%
prosody: pitch	72.6%
prosody: last 200 ms	70.3%
prosody: all	74.5%
all feature sets combined	79.7%

Table 1: Performance accuracy of each feature set in predicting question-bearing turns in the human-human ITSpoke corpus.

Percentage	Feature
1.3%	pre-test score
1.3%	ratio of rising slope of last 200 ms
1.3%	maximum pitch of entire turn
1.3%	cummulative pause duration
1.2%	regression slope of last 200 ms
1.1%	regression slope of entire turn
1.1%	mean pitch of entire turn
1.0%	mean loudness of last 200 ms
1.0%	maximum loudness of entire turn
1.0%	point slope of last 200 ms

Table 2: The most-used features in the learned decision tree from the machine learning experiment using all features.

turn, and the tutor dialogue act immediately preceding the student turn (also hand-labeled). The possible turn correctness labels are: *fully, partially, none, not applicable*. Tutor dialogue acts comprise: *short answer question, long answer question, deep answer question, positive feedback, negative feedback, restatement, recap, request, bottom out, hint, expansion, non-substantive*[2].

5. Machine Learning Experiments

In our corpus of tutorial dialogues most student turns do not contain questions. Excluding student turns that function only to maintain discourse flow, such as back-channels (*e.g.,* 'uh huh'), non-question-bearing student turns outnumber question-bearing turns nearly 2.5 to 1. In order to learn meaningful cues to questions and avoid a machine learning solution that favors non-question-bearing turns *a priori*, we down-sampled the latter turns from each student to match the number of question-bearing turns for that student. Thus the majority class baseline was 50%.

We conducted nine classification experiments to evaluate the usefulness of different types of features described above in predicting question-bearing turns, as well as to examine the predictive power of all feature sets combined. A final experiment was also conducted using all prosodic features calculated over only the last 200 milliseconds of each student turn. Each classification experiment used the WEKA machine learning environment [12]. While we experimented with several machine learning algorithms, including decision trees, rule induction, and support vector machines, we present results for the decision tree learner C4.5 boosted with the meta learning algorithm ADABOOST [13], which provided the best results. Performance accuracy for each experiment was averaged after running 5-fold cross validation.

6. Results

Our findings indicate prediction accuracy of student question-bearing turns in the human-human ITSpoke data of 79.7% using all features in aggregation. Furthermore, the precision, recall, and F-measure using all features are each 0.8, showing that this performance accuracy is robust.

Table 1 shows the performance accuracy of each feature set described in Section 5 in isolation. Here we see that the least predictive feature sets are rhythmic (52.6%) and student and task dependent (56.1%). The most predictive feature set comprises all prosodic information (74.5%), though it appears that the most significant contributor to this set is the prosodic information of the last 200 milliseconds of student turns (70.3%). The performance accuracies of the remaining feature sets fall somewhere in between.

The individual features with highest information gain are all prosodic: the pitch slope of the last 200 milliseconds (0.16), the maximum pitch of the entire turn (0.12), the pitch slope of the entire turn (0.09), and the mean pitch of the entire turn (0.08). However, non-prosodic features are also somewhat informative. The most informative syntactic features are the following: personal pronoun followed by a verb (0.04), interjection (0.03), determiner followed by a noun (0.02), *wh*-pronoun (0.02), and modal auxiliary followed by a personal pronoun (0.02). The most informative lexical ngrams are the following: 'yes' (0.03), 'right' (0.02), 'what' (0.02), 'I' (0.02), 'that' (0.02), and 'you' (0.02).

Table 2 lists the most frequently used features in the learned decision tree for the experiment in which all features were used together. The most-used individual features, each accounting for 1.3% of all decisions, are student pre-test score, ratio of rising slope of last 200 ms, maximum pitch of entire turn, and cumulative pause duration.

7. Discussion

From these experiments, we see that prosodic information is clearly the most useful indicator of the presence of a student question-bearing turn. Of these features, pitch information – especially pitch slope at the end of the turn – is the most useful. This is not in itself surprising, if most of these questions are rising [7]. However, this finding is encouraging nonetheless for spoken ITSs, since it suggests that, even though we are examining full student turns rather than hand-segmented questions, we can still identify these question-bearing turns by their prosody. Our broader analysis of question-bearing turns does indicate that, when students ask a question, it is usually the primary function of the turn.

Although turn-final pitch slope appears to be the most useful feature for predicting question-bearing turns, the fact that all features combined perform better than the prosodic feature set alone indicates that other features also contribute. Both lexical and POS ngrams improve overall performance, although they are somewhat redundant. For example, both the word 'what' and the part of speech that groups *wh*-pronouns are informative features. However, a few lexical and syntactic features stand apart. Interjections – words such as 'um', 'hm', 'alright', and 'sorry' – are the second most informative part of speech in detecting question-bearing

[2]For further explanation of ITSpoke dialogue act labels, we refer the reader to [11].

turns. With respect to lexical information, it is notable that lexical-pragmatic words are more informative than lexical-syntactic ones. For example, words such as 'yes' and 'right' have slightly higher information gain than does the word 'what'. The fact that both types of information are present in questions does not contradict previous findings, as described in Section 1, and though we can't be certain that our findings necessarily hold for all questions in general, it is very intriguing that lexical-pragmatic information appears to be just as useful as lexical-syntactic information for the identification of question-bearing turns.

What role do the remaining features play? At first glance, it appears that student and task dependent features contribute nothing to the prediction of question-bearing turns. However, the frequent appearance of student pre-test scores in the decision tree is suggestive. Although in isolation it provides no information gain, it may be that a pre-test score helps to contextualize other features. We notice in our corpus that as student pre-test scores increase, the ratio of *yes-no* questions (*e.g.,* "Is it gravity?") decreases whereas the ratio of *yes-no* tag questions (*e.g.,* "That would be gravity, right?") increases. An analogous pattern may exist for question-bearing turns as well. For example, phrase-final rising $f0$ may identify a question more accurately for students with low pre-test scores. Examination of this hypothesis is one of our future goals.

Though pitch information is most useful in this experiment, it is an open question whether this will also occur when students interact with an automated tutor. In initial and informal investigation of ITSpoke data collected of students interacting with such an automated tutor, we notice that rising pitch is indeed often apparent, possibly even more so than in the human-human environment. This is a second question we will test in future experiments.

8. Conclusion

Detecting whether or not a student turn contains a question is clearly useful for ITSs, since successful systems must meet the social expectations of their users. When one party in human-human conversation asks a question, the conversational partner normally responds. A first goal of our research has been to determine whether such questions are detectable via automatic means. Our results indicate that we can indeed recognize question-bearing turns with considerable accuracy (79.7%).

However, not all questions expect the same type of response. Some questions seek novel information while others seek clarification or acknowledgment. In order to meet student needs then, ITSs – and spoken dialogue systems in general – must not only be able to identify the presence of a question in a turn, but they must be able to determine its function. We have begun preliminary work to address this concern. To this end, the corpus has been hand-labeled for question function. Using the same features we have outlined above, we have run initial machine learning experiments showing us that, *given that we know a student turn bears a question*, we can predict the function of this question with about 75% accuracy. The most important feature for this task appears to be pragmatic: the previous tutor dialogue act, which, of course, will be available to the ITS. Other informative features appear to be lexical and syntactic information. Prosodic information appears to be least useful in this regard. Our future work will explore these issues in more detail.

9. Acknowledgments

This research was supported in part by NSF grant IIS-0328295. We thank Diane Litman, Kate Forbes-Riley, Mihai Rotaru, and Scott Silliman from the Research and Development Center at the University of Pittsburgh for data collection, annotation, and discussion.

10. References

[1] Arthur C. Graesser and Natalie K. Person, "Question asking during tutoring," *American Educational Research Journal*, vol. 31, no. 1, pp. 104–137, Spring 1994.

[2] Lisa Anthony, Albert Corbett, Angela Z. Wagner, Scott M. Stevens, and Kenneth R. Koedinger, "Student question-asking patterns in an intelligent algebra tutor," in *Proceedings of the International Conference on Intelligent Tutoring Systems*, Maceio, Brazil, 2004, pp. 455–467.

[3] Diane Litman and Scott Silliman, "Itspoke: An intelligent tutoring spoken dialogue system," in *Proceedings of the 4th Meeting of HLT/NAACL* (Companion Proceedings), Boston, MA, May 2004.

[4] Ronald Geluykens, "Intonation and speech act type. an experimental approach to rising intonation in queclaratives," *Journal of Pragmatics*, vol. 11, pp. 483–494, 1987.

[5] Robbert-Jan Beun, "The recognition of Dutch declarative questions," *Journal of Pragmatics*, , no. 14, pp. 39–56, 1990.

[6] Marie Šafářová and Marc Swerts, "On recognition of declarative questions in English," in *Proceedings of Speech Prosody*, Nara, Japan, March 2004.

[7] Janet B. Pierrehumbert and Julia Hirschberg, "The meaning of intonation contours in the interpretation of discourse," in *Intentions in Communication*, P. R. Cohen, J. Morgan, and M. E. Pollack, Eds., pp. 271–311. MIT Press, 1990.

[8] E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Vav Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech," *Language and Speech*, vol. 41, no. 3-4, pp. 439–487, 1998.

[9] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[10] Jackson Liscombe, Julia Hirschberg, and Jennifer Venditti, "Detecting certainness in spoken tutorial dialogues," in *Proceedings of Interspeech*, Lisbon, Portugal, 2005.

[11] Kate Forbes-Riley, Diane Litman, Alison Huettner, and Arthur Ward, "Dialogue-learning correlations in spoken dialogue tutoring," in *Proceedings 12th International Conference on Artificial Intelligence in Education (AIED 2005)*, Amsterdam, July 2005.

[12] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with Java implementations," in *ICONIP/ANZIIS/ANNES'99*, Dunedin, New Zealand, November 1999, pp. 192–196.

[13] Yoav Freund and Robert E. Schapire, "A short introduction to boosting," *Journal of the Japanese Society for Artificial Intelligence (JSAI)*, vol. 14, no. 5, pp. 771–780, 1999.