Question Classification in an Epistemic Game

Haiying Li[1], Borhan Samei[1], Andrew M. Olney[1], Arthur C. Graesser[1], and David W. Shaffer[2]

[1]University of Memphis, Institute for Intelligent Systems, Memphis, USA
{hli5,bsamei,aolney,graesser}@memphis.edu
[2]University of Wisconsin-Madison, Departments of Educational Psychology and Curriculum and Instruction, Madison, USA
{dws}@education.wisc.edu

Abstract. During collaborative problem solving in the epistemic game, *Land Science*, players asked questions when they were uncertain about concepts they were learning or the tasks they had been assigned. In our scenario, one mentor handles questions from several groups of students via computer-mediated chats. Because the mentor must track the learning progress of multiple groups of students simultaneously, it is common for the mentor to be unable to answer all student questions during the game. This paper presents an automated question classifier designed to augment the human mentor and enhance the ability to answer student questions. The question classification model used 30 linguistic features consisting of keyword lists and part of speech tags. The model was trained and tested with J48 Decision Trees. The results showed a good performance on question classification with the selected features. The question classifier will ultimately be used for the development of the automated question answering in the epistemic game.

Keywords: question classifier · collaborative problem solving epistemic game

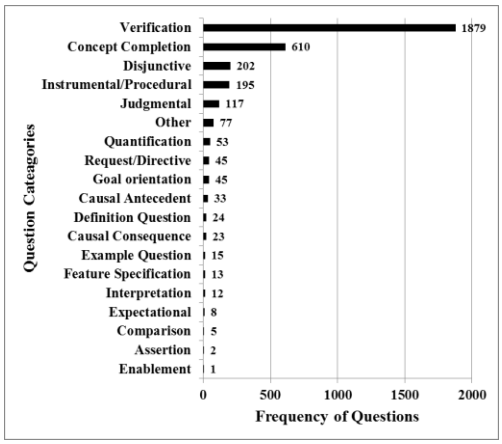
1 Instruction

Collaborative problem solving is a critical and necessary skill [1] because successful organizations need people who effectively work in teams. High quality collaborative work requires such skills as organizing a team, establishing a common ground and vision, identifying shared goals, assigning tasks, tracking progress, building consensus, troubleshooting potential problems, managing conflict, and maintaining communication [2,3]. Murray, Stephens et al. [4] proposed that the underlying skills mentioned above are necessary to address heterogeneous and complex social issues, and they called these skills *social deliberative skills* (SD-Skills). SD-Skills consist of a subset of higher order skills, including social metacognition [5], reflective judgment and epistemic skills [6], social perspective taking and empathy [7], and perspective seeking and question asking [8]. Question asking, one of higher order skills in SD-Skills [4], is indispensable during collaborative problem solving, especially in epistemic games. Epistemic games used rules to manipulate "abstract forms of knowledge or schemata appropriate to a discipline" (p. 228) [15]. In the game, student will fre-

quently ask questions to understand the ideas, opinions and perspectives of others, particularly when he/she is stuck, confused or uncertain [9]. For example, in *Land Science* [16] game, a mentor handles more than 10 students, who are in 3 or 4 groups. Such a human mentor is easily overwhelmed by student questions, which leaves many questions unaddressed. Therefore, the instant identification of questions and the immediate response to questions seem critical for the success of game-based learning environments when mentors are shared across groups.

Automated question classification is a top priority for dialogue systems in the epistemic games. Prior research on automated question classification focused on either the feature selections in dialog systems [10,11] or question taxonomy in tutoring systems [12]. However, less research has focused on developing automated question classification using question taxonomies that are well-aligned with epistemic games [13]. The present study developed an automated question classification using *Graesser-Person taxonomy* [12] with slight modifications (see below, this section). The question categories in the Graesser-Person taxonomy were defined according to the content of the information sought in empirical analyses of tutoring data [12]. The taxonomy has 18 question categories (see Figure 1), which were defined on the basis of the content of the questions and expected answers rather than simple signal words (e.g., "who" and "what"). The Graesser-Person taxonomy has been successfully applied to dialogues with intelligent tutoring systems [15]. We added an additional category (*Other)* for chat contributions containing only a question mark, a common turn in online chats. This question taxonomy was used to code the questions in the epistemic game, *Land Science*.

Fig. 1. Distribution of Question Categories of the Mentor and Students in Land Science Game



Land Science is an interactive urban-planning simulation with collaborative problem solving in an online game environment [15,16]. Players are assigned an in-game internship in which they act as land planners in a virtual city. They are guided through the game by a human mentor. Players are split into three or four groups in each game to accomplish the required tasks. An individual player communicates with the mentor and 3-5 other players within the same group through text chats when they have prob-

lems, complete their tasks, or send their products. They are not allowed to communicate verbally even though sometimes they are in the physical presence of other students. From the approximate 10-hour game, students learn to think like urban planners, represent the stakeholders to make the land use decisions, and balance the social and environmental issues. Thus, communities can meet their own needs and serve the public interests.

2 Method

Ninety-one high school and middle school students participated in seven Land Science games in two formats: on-site game during the vacation week or in school, and in distance but with one on-site meeting. The total 26,148 text chats were generated from seven games, among which 3,359 were questions, with 1,418 questions asked by the mentor and 1,941 by students. We used the 3359 questions to train and test our model.

Two researchers classified 3,359 students' questions according to the 19 question taxonomy. They first rated 68 questions, and kappa agreement scores averaged .50. After discussing disagreements, they started coding the entire corpus. The averaged kappa was .78 across all the categories. Thus, one of the researchers continued coded the mentor's questions. The model was trained and tested with this researcher's coding results.

Our feature set consists of a total number of 30 features [13]. The features represent presence or absence (binary) in the utterances, including (1) parts of speech (e.g., Determiner, Noun, Pronoun, Adjective, Adverb, and Verb), (2) word lists of certain word category (e.g., Do/Have, Be, Modal, and *wh*-question words), and (3) some specific keywords corresponding to the particular question category [12], such as casual consequent words (e.g. "results," "effects," etc.), procedural words (e.g. "plan," "scheme," "design," etc.), and others like Feature specification, Negation, Meta-communication, Metacognition, Comparison, Goal Orientation, Judgmental Definition, Enablement, Interpretation, Example, Quantification, Casual Antecedent, Disjunction. We also checked the binary attributes of particular words such as "happen," "no," and "yes" [see 14 for detail].

Besides the binary attributes, we also looked at the position of the captured attribute in the utterance, such as whether the word that was being captured in the beginning (the first word), in the middle (the word between the first and the last), or in the end (the last word). If the utterance only had one word, we coded as beginning; if the utterance had two words, we coded the first as beginning, and the second as end.

The present study employed a J48 decision tree to train and test the model of question classification using WEKA [17].

3 Results and Discussion

We built a J48 decision tree model on the human annotated data set, and evaluated the model with 10-fold cross validation. J48 is an implementation of C4.5 algorithm,

which is used to generate a decision tree [18]. The distribution of question categories, however, was skewed, and some categories included few instances, such as Enablement, Assertion, Comparison, Expectational, Interpretation, Feature Specification, and Example in the data set (see Figure 1). We combined the less frequent categories into one category called "Other" and evaluated our model on both the original distribution with all the categories and the data with less frequent categories combined into one category. Table 1 showed the performance of these two models.

Table 1. Performance of J48 Models with All and Combined Categories

	All Categories	Frequent Categories
Accuracy (%)	79.56	79.92
Kappa	0.66	0.67

As seen in Table 1, the performance of J48 models was high no matter which data set was used to train and test the model, above 79%. Using ZeroR algorithm as a baseline, the models performed with an accuracy of 55.8%. The machine learning question classification yielded similar results to the two human raters. The findings imply that the selected features for the question classification were sufficient to classify the questions even though the distribution of categories was not uniform.

The models were designed for the Land Science game, and the less frequent categories were the types of questions that were unlikely to emerge in this context. However, the less frequent question categories are likely to be used in production rules to generate automated response. Therefore, Table 2 showed the precision, recall and F-Measure based on the full category model.

Table 2. Precision, Recall and F-Measure in the Full Category Model

Category	Precision	Recall	F-Measure
Verification	0.86	0.93	0.90
Disjunctive	0.79	0.93	0.86
Goal orientation	0.72	0.84	0.78
Concept completion	0.74	0.74	0.74
Judgmental	0.79	0.57	0.66
Instrumental/procedural	0.59	0.50	0.54
Other	0.76	0.40	0.53
Quantification	0.41	0.53	0.46
Causal antecedent	0.46	0.30	0.36
Causal consequence	0.33	0.26	0.29
Average	0.76	0.80	0.78

Nine categories showed the precision, recall and F-Measure were 0.00; therefore, these categories were not shown in the table. These nine categories were Assertion, Comparison, Definition, Enablement, Example, Expectation, Feature Specification, Interpretation and Request/Directive. Results indicated that such categories as verification and disjunctive questions had extremely high F-Measure scores, above .86.

Similarly, two human raters had the highest kappas for these two categories, above .96. This suggests that the features that the model selected may be the same as the human raters. Other categories had high human inter-rater reliability, but low in our model, such as quantification (human kappa .96, but .46 in the model). The inconsistency may result from the different key linguistic features that the human raters used and our model used. Thus, in the future, we would select the better features for question categories with low accuracy.

4 Conclusion

In general, our model could provide a better performance of question classification. This allows us to develop the automated question answering during the collaborative problem solving for the epistemic games. For example, we could develop the production rules for the automated responses to high frequency questions. The answer could be triggered by a series of attributes, such as latent semantic space analysis, speech act categories, five SKIVE categories (e.g., epistemology, values, skills, identity and domain knowledge) [15], and/or regular expressions. With the automated question classification and question answering, an answer could be sent to the students immediately. This immediate answer may facilitate a better collaboration during the group learning and problem solving in the epistemic game.

Since some question categories had fewer frequencies, in the next phase, we would collect more data to improve our model. Specifically, we would build the question classification tool into an interface that would show the classified result for each question and allow a human expert to modify the results if the machine classified category is wrong. In this way, with more and more data in the database, we hope to obtain the more accurate results of the question classification.

The further analysis of the high-frequency questions in each stage of the game can help generate specific responses to each question. For example, in a given task, players may ask similar questions or for explanations or elaborations. Thus, certain questions could be addressed in advance.

5 Acknowledgement

This work was supported by the National Science Foundation (0918409) for the project of AutoMentor: Virtual mentoring and assessment in computer games for STEM learning. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these funding agencies, cooperating institutions, or other individuals.

6 References

1. OECD. PISA 2015 Collaborative Problem Solving Framework. OECD Publishing (2013)

2. Dillenbourg, P., Traum, D.: Sharing Solutions: Persistence and Grounding in Multimodal Collaborative Problem Solving. Journal of the Learning Science, 15, 121–151 (2006)
3. Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., Warner, N.: Toward an Understanding of Macrocognition in Teams: Predicting Processes in Complex Collaborative Contexts. Human Factors: The Journal of the Human Factors and Ergonomics Society, 52, 203–224 (2010)
4. Murray, T., Stephens, A.L., Woolf, B.P., Wing, L., Xu, X., Shrikant, N.: Supporting Social Deliberative Skills Online: The Effects of Reflective Scaffolding Tools. In: 5th International Conference on Online Communities and Social Computing eSociety at HCII 2013. Las Vegas (2013)
5. Lin, X., Sullivan, F.: Computer Contexts for Supporting Metacognitive Learning. In: Voogt, J., Knezek, G. (eds.) International Handbook of Information Technology in Primary and Secondary Education, 281–298. Springer Science+Business Media, LLC. (2008)
6. Kuhn, D.: Metacognitive Development. Current Directions in Psychological Science, 9, 178-181 (2000)
7. Suthers, D. D., Desiato, C.: Exposing Chat Features Through Analysis of Uptake Between Contributions. In: 45th Hawaii International Conference on the System Sciences, pp. 3368-3377, IEEE Press, New York (2012)
8. Graesser, A.C, Rus, V., Cai, Z.: Question Classification Schemes. In: Proceedings of the Workshop on Question Generation (2008)
9. Li, H., Duan, Y., Clewley, D., Morgan, B., Graesser, A. C., Shaffter, D. W., Saucerman, J.: Question Asking During Collaborative Problem Solving in an Online Game Environment. In: The 12th International Conference on Intelligent Tutoring Systems. Springer-Verlag (2014)
10. Li, X., Roth, D.: Learning Question Classifiers. In: The 19th International Conference on Computational Linguistics, Vol. 1, pp. 1–7 (2002)
11. Huang, Z., Thint, M., Qin, Z.: Question Classification Using Head Words and Their Hypernyms. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 927-936 (2008)
12. Graesser, A.C., Person, N.K.: Question Asking During Tutoring. American Educational Research Journal. 31, 104–137 (1994)
13. Olney, A.M., Louwerse, M., Mathews, E.C., Marineau, J., Mitchell, H.H., Graesser, A.C.: Utterance Classification in AutoTutor. Human Language Technology – North American Chapter of the Association for Computational Linguistics 1–8 (2003)
14. Olney, A. M. Gnututor: An Open Source Intelligent Tutoring. System Based on AutoTutor. In: Proceedings of the 2009 AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems, pp. 70–75. Washington, DC: AAAI Press (2009)
15. Shaffer, D.W.: Epistemic Frames for Epistemic Games. Computers & Education. 46, 223-234 (2006)
16. Keshtkar, F., Burkett, C., Li, H., Graesser, A.C.: Using Data Mining Techniques to Detect the Personality of Players in an Educational Game. In: Pena-Ayala, A. (ed.) Educational Data Mining: Applications and Trends, pp. 125-150. New York: Springer (2014)
17. Witten, I.H., Frank, E., Trigg, L.E., Hall, M.A., Holmes, G., Cunningham, S.J. Weka: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers (2011)
18. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers (1993)