Assessing Question Quality Using NLP

Kristopher J. Kopp[1(✉)], Amy M. Johnson[1], Scott A. Crossley[2], and Danielle S. McNamara[1]

[1] Department of Psychology, Arizona State University, Tempe, AZ 85287, USA
{Kjkopp, amjohn43, Danielle.McNamara}@asu.edu
[2] Department of Applied Linguistics/ESL, Georgia State University, 25 Park Place, Atlanta, GA 30303, USA
scrossley@gsu.edu

Abstract. An NLP algorithm was developed to assess question quality to inform feedback on questions generated by students within iSTART (an intelligent tutoring system that teaches reading strategies). A corpus of 4575 questions was coded using a four-level taxonomy. NLP indices were calculated for each question and machine learning was used to predict question quality. NLP indices related to lexical sophistication modestly predicted question type. Accuracies improved when predicting two levels (shallow versus deep).

Keywords: Intelligent tutoring systems · Artificial intelligence · Natural language processing · Educational technology design · Question classification

1 Introduction

iSTART (Interactive Strategy Training for Active Reading and Thinking) is an ITS that provides instruction on self-explanation strategies and generative strategy practice with immediate feedback using natural language processing (NLP; [1]). Research indicates that iSTART improves learners' ability to construct quality self-explanations and increases reading comprehension [2]. Similar to self-explanation, question asking is an effective reading strategy and asking deep (i.e., questions that get at a deeper form of knowledge) rather than shallow questions during reading improves reading comprehension [3]. Researchers have created systems to *generate* questions for learners to answer during learning [4]. However, to our knowledge, no systems are available to *assess* the quality of questions that readers ask *during* reading.

Our goal is to create a mechanism to *provide feedback on questions* students ask while reading. The first step is to create an algorithm to classify deep vs. shallow level questions. Readers were explicitly instructed to ask questions and human coders applied a classification scheme modified from Graesser and Person question taxonomy [5] to classify the questions, producing the data for the development of the NLP algorithm described in this study.

E. André et al. (Eds.): AIED 2017, LNAI 10331, pp. 523–527, 2017.
DOI: 10.1007/978-3-319-61425-0_55

2 Method and Results

Two hundred thirty-three participants were recruited using the Amazon Mechanical Turk online research service. Participants read three short, simplified news articles that included three to seven pre-identified target sentences (164 total) for which participants produced questions. The dataset included 4,575 questions. Our coding scheme ranged from (1) very shallow to (4) very deep. Two trained researchers coded 60% of the data set each, with 20% overlap to establish the interrater reliability: $\text{kappa}_{(\text{linear weighted})} = .84$, $r = .67$, 82% exact agreement, and 92% adjacent agreement. Remaining differences between the coders were resolved.

Each question was run through a number of NLP tools including the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [6], the Tool for the Automatic Analysis of Cohesion (TAACO) [6] and the Constructed Response Analysis Tool (CRAT) [7]. We used the indices reported by the NLP tools to predict human scores (1 through 4) for the corpus of questions. Indices reported that lacked normal distributions were removed. A MANOVA was conducted using the NLP indices as dependent variables and the four categories of questions as independent variables. A DFA retained 28 variables (see Table 1 for the MANOVA results for variables retained in the DFA). The majority of these variables were related to lexical sophistication. The DFA correctly allocated 1904 of the 4575 questions in the total set, $\chi 2$ ($df = 9$, $n = 4575$) = 669.567, $p < .001$, accuracy = 41.6%. A leave-one-out cross-validation (LOOCV) analysis allocated 1834 of the 4575 texts, accuracy = 40.1%. Agreement between the human and the model produced a $\text{kappa}_{(\text{linear weighted})} = 0.21$. A similar analysis was conducted using two categories (shallow vs. deep) of questions as the independent variables. The DFA retained 14 variables (see Table 2) which were also mostly related to lexical sophistication. The DFA correctly allocated 2817 of the 4575 questions in the total set, $\chi 2$ (df = 1, $n = 4575$) = 245.063, $p < .001$, accuracy = 61.6%. A LOOCV analysis correctly allocated 2794 of the 4575 texts, accuracy = 61.1%. Agreement between the human and the model produced a $\text{kappa}_{(\text{linear weighted})} = 0.23$.

Table 1. List of indices and MANOVA results for four category analysis

Index	Greater at deeper level±	*F*	Partial N^2
Proportion of bigrams COCA (70,000 words)	Yes	39.247**	0.025
Average lexical decision accuracy	Yes	30.276**	0.019
Lemma TTR (content words)	Yes	24.608**	0.016
Log content word range COCA news	Yes	18.724**	0.012
Lemma overlap between question and text	Yes	18.945**	0.012
Mean combined concreteness score	Yes	18.634**	0.012
Word frequency: Thorndike Lorge (all words)	No	15.015**	0.010
Word frequency (log): BNC spoken content words	Yes	14.082**	0.009

(*continued*)

Table 1. (*continued*)

Index	Greater at deeper level±	*F*	Partial N^2
Word frequency (log): COCA spoken content words	Yes	10.377**	0.007
Proportion of bigrams COCA (80,000 words)	No	10.964**	0.007
Lemma TTR (news words)	Yes	6.965**	0.005
Proportion of bigrams COCA (50,000 words)	Yes	7.105**	0.005
Mean COCA bigram log frequency score	Yes	8.030**	0.005
Lemma TTR (COCA fiction)	Yes	7.967**	0.005
Standardized naming RT	No	5.911**	0.004
Bigram proportion score COCA (100,000 words)	Yes	6.374**	0.004
Lemmas TTR (magazine words)	Yes	5.944**	0.004
Semantic variability of contexts	Yes	6.352**	0.004
Lemma TTR (academic words)	No	3.949*	0.003
Lemma TTR (all words)	Yes	5.016*	0.003
Bigram proportion score BNC written words	Yes	4.044*	0.003
TTR for questions (content words)	Yes	4.098*	0.003
Academic bigram association strength (COCA)	Yes	5.085*	0.003
Bigram proportion score COCA (60,000 words)	No	3.436*	0.002
Lemma proportion COCA (fiction)	Yes	2.477*	0.002
Word frequency: COCA academic function words	No	3.094*	0.002
Word frequency: COCA spoken content words	Yes	2.967*	0.002
Log academic word range COCA (all words)	No	2.772*	0.002

*$p < .05$, **$p < .01$; TTR = type-token ratio

± Yes indicates average value for deep questions (level 3 and 4) was above the overall mean

Table 2. List of indices and MANOVA results for two category analysis

Index	Greater at deeper level±	*F*	Partial N^2
Average lexical decision accuracy	Yes	86.186**	0.018
Mean combined concreteness score	Yes	38.952**	0.008
Word frequency (log): BNC spoken (all words)	Yes	37.730**	0.008
Word frequency: Thorndike Lorge (all words)	No	31.145**	0.007
Word frequency (log): COCA spoken content words	Yes	24.156**	0.005
Word range COCA news (content words)	Yes	23.446**	0.005

(*continued*)

Table 2. (*continued*)

Index	Greater at deeper level±	*F*	Partial N^2
Content words TTR	Yes	21.350**	0.005
Semantic similarity across words in question	Yes	14.107**	0.003
Standardized naming reaction time across all participants for this word	No	14.244**	0.003
Word frequency (log): BNC (all words)	No	9.924*	0.002
Lemma TTR	Yes	7.480*	0.002
Lemma proportion COCA	No	7.397*	0.002
Bigram proportion score BNC written words	Yes	5.767*	0.001
Bigram proportion score COCA (60,000 words)	No	4.911*	0.001

*$p < .05$, **$p < .01$; TTR = type-token ratio

± Yes indicates average value for deep questions (level 3 and 4) was above the overall mean

3 Conclusions

The most predictive indices related to lexical sophistication and lexical and semantic overlap. Deeper level questions contained less sophisticated words and greater lexical and semantic overlap both within the question and with the text. They included words with higher accuracies on lexical decision tests, more frequent words, less specific words, and more concrete words. Deeper level questions contain words that are easier to process and more familiar allowing for better comprehension of the question. The current study takes strides towards automating classifications of question quality and contributes to the improvement of an existing ITS with the objective of enhancing reading comprehension for a wide range of readers [4]. Our hope is that future work that builds on this foundation will be beneficial to the development of other ITSs and a variety of computer-based learning environments.

Acknowledgments. This research was supported in part by the Institute for Educational Sciences (IES R305A130124) and the Office of Naval Research (ONR N00014-14-1-0343 and ONR N00014-17-1-2300). Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of IES or ONR.

References

1. McNamara, D.S.: Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. Discourse Process. **38**, 1–30 (2015)
2. McNamara, D.S., O'Reilly, T.P., Best, R.M., Ozuru, Y.: Improving adolescent learners' reading comprehension with iSTART. J. Educ. Comput. Res. **34**(2), 147–171 (2006)
3. Cerdán, R., Vidal-Abarca, E., Martínez, T., Gilabert, R., Gil, L.: Impact of question-answering tasks on search processes and reading comprehension. Learn. Instr. **19**(1), 13–27 (2009)

4. Graesser, A. C., Jackson, G.T., Mathews, E.C., Mitchell, H.H., Olney, A., Ventura, M., Chipman, P., Franceschetti, D., Hu, X., Louwerse, M.M., Person, N.K.: Why/autotutor: a test of learning gains from a physics tutor with natural language dialog. In: Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society, pp. 1–6 (2003)
5. Graesser, A.C., Person, N.K.: Question asking during tutoring. Am. Educ. Res. J. **31**(1), 104–137 (1994)
6. McNamara, D.S., Allen, L.K., Crossley, S.A., Dascalu, M., Perret, C.A.: Natural language processing and learning analytics. In: Siemens, G., Lang, C. (eds.) Handbook of Learning Analytics and Educational Data Mining (in press)
7. Crossley, S., Kyle, K., Davenport, J., McNamara, D.S.: Automatic assessment of constructed response data in a chemistry tutor. In: Barnes, T., Chi, M., Feng, M. (eds.) EDM 2016, pp. 336–340. International Educational Data Mining Society, Raleigh (2016)