Automatically Generating Discussion Questions

David Adamson[1], Divyanshu Bhartiya[2], Biman Gujral[3], Radhika Kedia[3], Ashudeep Singh[2], and Carolyn P. Rosé[1]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
{dadamson,cprose}@cs.cmu.edu
[2] IIT Kanpur, Uttar Pradesh, India
{divbhar,ashudeep}@iitk.ac.in
[3] DA-IICT, Gandhinagar, Gujarat, India
{biman_gujral,radhika_kedia}@daiict.ac.in

Abstract. Automatic question generation can support instruction and learning. However, work to date has produced mostly "shallow" questions that fall short of supporting deep learning and discussion. We propose an extension to a state-of-the-art question generation system that allows it to produce deep, subjective questions suitable for group discussion. We evaluate the questions generated by this system against a panel of experienced judges, and find that our approach fares significantly better than the baseline system.

Keywords: question generation, facilitation, subjectivity, computer-supported collaborative learning.

1 Introduction

Recent work, built on observations of expert classroom instruction, has advocated strategies for reading and knowledge-building that move beyond simple comprehension and into questioning and reasoning [1]. Additionally, deep reasoning questions in tutorial environments have been shown to be correlated with student learning [2,3,4]. Such questions offer opportunities for evaluation, multiple perspectives and opinions, and synthesis, corresponding to the higher ("deeper") levels of Bloom's taxonomy [5,6]. Effective automated support for deep learning should be able to produce contextually suitable deep questions . However, producing such questions automatically for a new text or domain has remained an unanswered challenge.

Automatic question generation can indeed support instruction and learning in computer-based settings [7,8,9]. Work to date has produced mostly shallow questions that are not intended to promote deep thought or discussion, or that depend on special features of a particular domain. In this paper, we propose an extension to a state-of-the-art question generation system [10], allowing it to produce deep, subjective questions suitable for group discussion.

In the section that follows, we review the literature and prior work in the areas of discussion-oriented learning, deep questions, and question generation.

K. Yacef et al. (Eds.): AIED 2013, LNAI 7926, pp. 81–90, 2013.

Sec. 3 describes our improvements to a baseline question generation system. Our evaluation method and analysis of results are described in Sec. 4 and 5, followed by discussion of the results and directions for future work.

2 Theoretical Framework

2.1 Discussion and Instruction

The literature of instructional practices has advocated strategies for reading and knowledge-building that move beyond comprehension into questioning and reasoning [1], including Questioning the Author [11], Reciprocal Teaching [12], and Collaborative Reasoning [13]. Drawing on observations and analysis of successful classroom instruction, Michaels, O'Connor, and Resnick describe a framework for academically productive talk [14,15] as a collection of discussion-facilitating questions that a teacher can use to promote rich student-centered conversation and collaboration. In a study with teachers employing similar strategies, students have shown steep growth in achievement on standardized math scores, transfer to reading test scores, and retention of transfer for up to 3 years [16]. The success of these approaches hinges on skillful use of elicitation strategies like deep questions to invite the kind of discussion that leads to learning.

2.2 Deep Discussion Questions

Deep questions, allowing for multiple perspectives and reflective answers, are associated with the "deep learning" levels of Bloom's taxonomy [5]. Past work has shown the use of deep-reasoning questions [6] to be significantly correlated with student learning. Several recent studies [2,3,4] have shown high-quality discussion questions and reflective knowledge-building activities to be associated with positive learning outcomes. Further work [17,18] argues that text comprehension can be significantly improved by replacing traditional IRE instruction (Intiation-Reply-Evaluation [19]) with discussion-based activities where students have opportunities to summarize, challenge, make predictions on questions that allow multiple answers, and respond to questions that require them to draw upon evidence from both the text and their own personal perspectives.

Questions containing a greater proportion of highly subjective words - that is, words expressing opinions and evaluations - allow for multiple answers and personal perspective [20]. Responses to such questions offer opportunities to be challenged and built upon. Work in this sphere has produced the SentiWordNet database [21], where word senses are associated with subjectivity scores. While measures of subjectivity have largely been used for opinion mining, the measure of the subjective potential of a question may serve as a convenient proxy for deepness. More objective questions may be "shallower" in that they may be answered simply and factually, whereas more subjective questions leave room for justification and opinion, aligning with the "deep" questions described above.

2.3 Question Generation

Recent work in question generation has focused on generating objectively answerable, fill-in-the-blank or multiple-choice questions [8,9,10]. These basic questions can be generated with some success, but do not necessarily promote discussion. Present methods prefer clear, answerable questions - but to promote discussion, multiple answers and perspectives must be possible.

Heilman [10] describes a system for producing reading questions from a text. Leveraging off-the-shelf NLP tools, each declarative sentence passes through a set of general-purpose structural transformations to produce a collection of candidate questions. These questions are then ranked by a model trained on human judgements, using lexical and structural features of the question. While this method creates reading comprehension questions that are reliably grammatical, they are recall-oriented, and are not intended as "deep questions".

Although there has been some preliminary work in generating more probing questions from a text, the questions thus generated are limited in scope and depend on particularities of the domain. For example, Wang [8] employs question templates specific to the domain of medical texts, and Liu [22] uses the structure of citations in an academic paper to produce questions that address argumentation style.

3 Generating Questions for Discussion

We describe changes to baseline sentence selection and question generation methods [10] in order to promote deeper, more subjective questions drawn from a text. Instead of over-generating questions from all sentences in the summary, we instead select a subset of sentences based on one of three models of sentence "relevance". In all cases, including our application of the baseline system, questions are generated from sentences selected from a human-generated summary of a longer "original" text. Two of our selection models also utilize information from the original text. A summary is a more suitable source for discussion questions because individual sentences are more likely to contain abstractions or synthesis of ideas from the original text. After generating questions from this reduced set of candidate sentences, we apply the baseline system's method for generating questions. We then apply a set of transformations to the result to produce a set of questions more suitable for discussion. A measure of question-level subjectivity allows us to anticipate these questions' potential for deeper reasoning and rich discussion.

3.1 Selecting Sentences

We examine three methods for sentence selection, drawing on the fields of text categorization [23,24], information retrieval [25], and summarization [26,27,28]. Each of these embodies a different intuition as to what makes a sentence particularly salient, as described in each subsection below.

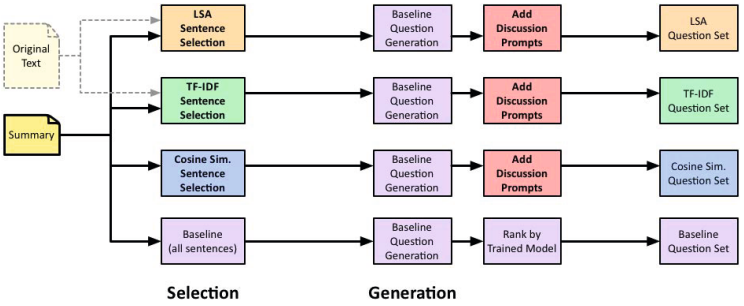

Fig. 1. System architecture, contrasted with the baseline system

Cosine Similarity. This method identifies sentence candidates from the summary using only the summary text. Considering each sentence as a "bag of words" vector, the cosine distance between two sentences is the angle between their word-vectors [24]. The smaller the cosine distance, the greater the similarity. Recognizing that the summary may highlight and build upon key concepts within its own structure, we calculate cosine similarity between each sentence of the summary text and the sentence preceding it. Sentences with high similarity to their immediate predecessors may be interpreted as marking an important concept, and as such are selected as candidates for question generation.

LSA Content Scores. Latent Semantic Analysis [23] is a technique designed to analyze the relationships between a set of documents (sentences, in our case) and the terms they contain. Each sentence is represented as an N-dimensional vector, where each dimension's value roughly corresponds to a sentence's weight for a "topic" in the original document set. We reduce the term-sentence matrix of the original text to an N-dimensional LSA space (N=5 in our case, although we did not tune this value), and also transform each sentence from the summary into its own vector in this space. Our goal, comparable to a text summarization task [26,27], is to select sentences most representative of each dimension. We select those sentences with the highest weight in each of the "topic" dimensions, producing N sets of candidate sentences from the summary.

TF-IDF Uniqueness. Term Frequency-Inverse Document Frequency is a metric used in information retrieval to measure the importance of a word [25]. In a given document (a candidate sentence in the summary text), the TF-IDF score of a word is the count of its occurrences in that document, multiplied by a factor (the inverse document frequency) that discounts its appearances in the entire corpus (in our case, the original text). Here TF-IDF is being applied as a measure of uniqueness, preferring those sentences in the summary with higher averaged per-word TF-IDF scores. Sentences from the summary with a high TF-IDF score contain a greater proportion of "rare" words relative to the source text, and thus may contain new ideas that are not literally present in the original.

3.2 Transforming and Ranking Questions

We further transform some of the more factoid-like questions generated by the baseline system into more subjective questions. When a simple yes-or-no question is extracted by the original system, we transform it into a "why" question, for example "(Why) does psychological manipulation prevent the common animals from doubting the pigs' abilities?". Other factoid questions are transformed by prompting for justification or elaboration, for example the question "What was inscribed on the side of the barn?" is appended with "Discuss in detail." While these transformations are nearly trivial to apply, they do transfer the responsibility of evaluation from the asker to the answerer. Such simple moves can empower students and promote productive discussion [14].

To rank the questions on the basis of abstraction and ability to trigger discussion, we calculate a *subjectivity* score for each question. Subjectivity may stand as a measure for "deepness", as described in Section 2.2. Question subjectivity is taken as an average of the subjectivity values of each word in the sentence, as given by SentiWordNet [21]. SentiWordNet is a database of words-senses, differentiated by part-of-speech, with subjectivity scores assigned to each. In the case where a word has more than one sense for a given part of speech, we take the average of its senses' subjectivity values.

4 Evaluation

We generated 50 questions using the baseline method [10] from an analysis and summary [29] of George Orwell's *Animal Farm* [30]. These were the top 50 questions as ranked by the system's trained model. We also generated questions using the methods described in this paper, and selected 50 of these at random. For discussion of texts in literature courses, we can rely on the bounty of existing human-authored summaries and analyses (like SparkNotes) to draw our questions from, although in future work we would like to incorporate an automatic summarization method.

A group of four teachers served as judges and evaluated this combined set of questions. Each judge received the questions in a random order. For each generated question, the judges rated their agreement with six statements about the question on a Likert scale, from 1-7. The first three of these statements

Table 1. Question evaluation dimensions

1 This question lends itself to multiple answers.
2 Answering this question could engage a student's personal values or perspective.
3 This question would be valuable for stimulating discussion among students.
4 This question touches upon important themes from the story.
5 This question is comprehensible.
6 This question is grammatical.

(shown in Table 1) correspond to Bloom's [5] and Graesser's [6] descriptions of the sort of deep-level questions that have been shown to be effective in tutorial settings [2]. The fourth statement probes the suitability of the question content. The last two dimensions are indicators of quality of the question's form. While none of these dimensions is inherently more important than another, a method for generating high-quality discussion questions should receive high scores in all dimensions.

5 Results and Analysis

In order to evaluate the relative quality of questions generated with our approach in comparison with the baseline method, as well as to compare among different selection criteria used by our method, we used an ANCOVA model for each of the six dimensions evaluated by the judges. For each dimension, the dependent measure was the rating assigned by the judge for that dimension. The independent variable was binary, indicating whether the rating was assigned to a question generated with the baseline approach or one of the experimental approaches. In order to differentiate among the three selection methods used by the experimental approach, we included a three-way categorical variable nested within the main independent variable. This allows us to test simultaneously whether the experimental approach is better than the control condition, and whether there are differences between the experimental approach's selection methods. In order to control for systematic differences between judges, we included a categorical control variable indicating which of the four judges assigned the score. A summary of the human ratings is displayed in Fig. 2. The Subjectivity score was used as a covariate in order to evaluate the effect of using Subjectivity as part of a selection criteria for discussion questions.

Multiple Answers. In terms of potential for eliciting multiple student answers, the judges rated the set of experimental approaches significantly better than the baseline approach $F(1, 288) = 12.3, p < .0005$, effect size .64 s.d. There were also significant differences between experimental approaches $F(2, 288) = 3.74, p < .05$ such that LSA and Cosine were significantly better than TF-IDF,

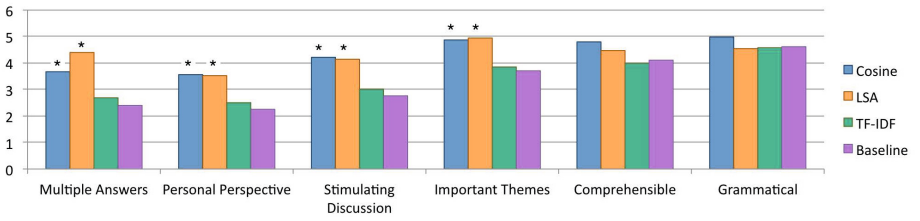

Fig. 2. Average rating per selection method for each dimension. A star ($\star$) indicates values which are significantly better than the baseline.

and TF-IDF was not significantly different from baseline. There was a marginal positive correlation between Subjectivity and the dependent measure ($p = .1$), indicating some support for using a subjectivity score as part of a selection method for discussion questions.

Personal Perspective. The results for a questions' potential to engage personal perspective were consistent with those for eliciting multiple answers. The judges rated the set of experimental approaches significantly better than the baseline $F(1, 288) = 8.2, p < .005$, effect size .39 s.d. There were also significant differences between experimental approaches $F(2, 288) = 3.02, p < .05$ such that LSA and Cosine were significantly better than TF-IDF, and TF-IDF was not significantly different from baseline. For this dimension, there was a significant positive correlation between Subjectivity and the dependent measure ($R = .13, p < .05$), suggesting that questions scored as more subjective offer students more opportunity to express their personal perspective.

Stimulating Discussion. Again, results for potential to stimulate discussion were the same. The judges rated the set of experimental approaches significantly better than the baseline approach $F(1, 288) = 9.6, p < .005$, effect size .43 s.d. There were also significant differences between experimental approaches $F(2, 288) = 3.28, p < .05$ such that LSA and Cosine were significantly better than TF-IDF, and TF-IDF was not significantly different from baseline. Again, there was a significant positive correlation between Subjectivity and the dependent measure ($R = .11, p < .05$), suggesting that questions that are scored as more subjective are rated as more stimulating for discussion.

Important Themes. Results for capturing important themes were distinct, although they still favored the experimental approach. This time, Subjectivity had no effect, and there were no significant distinctions among experimental approaches. However, there was a significant advantage attributed to the experimental approaches as a set over that of the baseline approach, $F(1, 288) = 7.05, p < .05$, effect size .37 s.d.

Comprehensibility. In terms of comprehensibility, the experimental approaches as a set were rated as marginally better than the baseline approach $F(1, 288) = 3.22, p < .1$. There were no differences among experimental approaches. And, in contrast to the other metrics, Subjectivity had a negative correlation with comprehensibility ($R = .19, p < .0005$).

Grammaticality. In terms of grammaticality, there were no significant differences among approaches. However, similar to the comprehensibility rating, Subjectivity had a negative correlation with grammaticality ($R = .17, p < .005$).

6 Discussion and Future Work

Broadly, we find that our method for generating questions from a summary text significantly outperforms the baseline system on those dimensions related to their suitability for classroom discussion. Table 2 illustrates some representative questions and scores produced by the three selection methods of our approach, as well as the baseline system.

Table 2. Representative questions generated by our system and the baseline on each of the 6 dimensions presented in Sec. 4 *Subj.* is determined as per Sec. 3.2

Selection Method	Question	Subj. Score	1 MA	2 PP	3 SD	4 ITT	5 Com	6 Gra
Cosine Sim.	Why does psychological manipulation unite the animals against a supposed enemy ?	0.26	5.5	5.75	6.25	6.25	6.5	6.5
TF-IDF	Whose idealism leads to his downfall?	0.29	3.25	2.75	2.75	4.5	7	7
LSA	What does the increasing frequency of the rituals bespeak? Discuss in detail.	0.18	5.5	4.5	5.25	5.5	5.25	4
Baseline	Who gathers the animals of the Manor Farm for a meeting in the big barn?	0.09	1	1.25	1.25	2.75	7	7

We note that although the questions generated from sentences selected by the LSA and by Cosine Similarity methods are rated nearly identically in each dimension, the set of questions they generate are quite different from each other. The Cosine Similarity selection method relies on the structure of the summary to highlight concepts worthy of discussion, and in so doing captures repeating elements - not just story words like “animals” and“windmill”, but more abstract themes developed in the summary. The LSA method, by contrast, selects a set of sentences from the summary that most strongly echo the latent “topics” of the original text, which can include both chronological associations (the character of Snowball is much more prevalent in the early story) and repeated themes (“Animalism”, “pigs”, “men”, “power”, and “equal” are favored by a single LSA-space dimension, highlighting the recurring contrast between the animals’ society and the humans’). The TF-IDF selection method favors sentences that are unique in comparison to the original document, which could potentially highlight those sentences which synthesize or abstract ideas not made explicit in the story. In practice however, the questions produced from the sentences selected by this method are short and specific, picking up on details in individual sentences that have less relationship to the story as a whole. It is thus unsurprising that this selection method fares no better than the baseline.

To evaluate the suitability of discussion questions in an educational setting, a prototype conversational agent has been implemented. Adapting the "revoicing" behavior described by Dyke and colleagues [31], the agent facilitates discussion on a given text by prompting the group with discussion questions (drawn from any one of the methods described in this paper) that are similar to statements made by the students (the details of this system is beyond the scope of this paper). In addition to piloting this system with students, future work might explore ways to scaffold a discussion session, perhaps by starting with more concrete questions, with lower subjectivity scores, and transition to deeper, more subjective questions as the discussion progressed.

References

1. Palincsar, A.: Collaborative approaches to comprehension instruction. In: Rethinking Reading Comprehension, pp. 99–114 (2003)
2. Graesser, A., Person, N.: Question asking during tutoring. American Educational Research Journal 31(1), 104–137 (1994)
3. Craig, S., Sullins, J., Witherspoon, A., Gholson, B.: The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. Cognition and Instruction 24(4), 565–591 (2006)
4. Roscoe, R., Chi, M.: Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors explanations and questions. Review of Educational Research 77(4), 534–574 (2007)
5. Bloom, B., Engelhart, M., Furst, E., Hill, W., Krathwohl, D.: Taxonomy of educational objectives: Handbook i: Cognitive domain, vol. 19, p. 56. David McKay, New York (1956)
6. Graesser, A., Rus, V., Cai, Z.: Question classification schemes. In: Proc. of the Workshop on Question Generation (2008)
7. Brown, J., Frishkoff, G., Eskenazi, M.: Automatic question generation for vocabulary assessment. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 819–826. Association for Computational Linguistics (2005)
8. Wang, W.-M., Hao, T., Liu, W.: Automatic question generation for learning evaluation in medicine. In: Leung, H., Li, F., Lau, R., Li, Q. (eds.) ICWL 2007. LNCS, vol. 4823, pp. 242–251. Springer, Heidelberg (2008)
9. Agarwal, M., Mannem, P.: Automatic gap-fill question generation from text books. In: Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 56–64. Association for Computational Linguistics (2011)
10. Heilman, M., Smith, N.: Question generation via overgenerating transformations and ranking. Technical report, DTIC Document (2009)
11. Beck, I., et al.: Questioning the Author: An Approach for Enhancing Student Engagement with Text. ERIC (1997)
12. Palinscar, A., Brown, A.: Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. Cognition and Instruction 1(2), 117–175 (1984)
13. Chinn, C., Anderson, R., Waggoner, M.: Patterns of discourse in two kinds of literature discussion. Reading Research Quarterly 36(4), 378–411 (2001)
14. Michaels, S., O'Connor, C., Resnick, L.: Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. Studies in Philosophy and Education 27(4), 283–297 (2008)

15. Resnick, L., Michaels, S., O'Connor, C.: How (well structured) talk builds the mind. In: Innovations in Educational Psychology: Perspectives on Learning, Teaching and Human Development, pp. 163–194 (2010)
16. Adey, P., Shayer, M.: An exploration of long-term far-transfer effects following an extended intervention program in the high school science curriculum. Cognition and Instruction 11(1), 1–29 (1993)
17. Langer, J.: Envisioning Literature: Literary Understanding and Literature Instruction. Language and Literacy Series. ERIC (1995)
18. Applebee, A., Langer, J., Nystrand, M., Gamoran, A.: Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. American Educational Research Journal 40(3), 685–730 (2003)
19. Cazden, C.B.: Classroom Discourse: The Language of Teaching and Learning. Heinemann, Portsmouth (1988)
20. Wiebe, J., Mihalcea, R.: Word sense and subjectivity. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, pp. 1065–1072. Association for Computational Linguistics (2006)
21. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta. European Language Resources Association, ELRA (May 2010)
22. Liu, M., Calvo, R.A., Rus, V.: Automatic question generation for literature review writing support. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 45–54. Springer, Heidelberg (2010)
23. Dumais, S.T.: Latent semantic analysis. Annual Review of Information Science and Technology 38(1), 188–230 (2004)
24. Huang, A.: Similarity measures for text document clustering. In: Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC 2008), Christchurch, New Zealand, pp. 49–56 (2008)
25. Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting tf-idf term weights as making relevance decisions. ACM Trans. Inf. Syst. 26(3), 13:1–13:37 (2008)
26. Wang, D., Zhu, S., Li, T., Gong, Y.: Multi-document summarization using sentence-based topic models. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 297–300. Association for Computational Linguistics (2009)
27. Dredze, M., Wallach, H., Puller, D., Pereira, F.: Generating summary keywords for emails using topics. In: Proceedings of the 13th International Conference on Intelligent User Interfaces, pp. 199–206. ACM (2008)
28. Hu, M., Sun, A., Lim, E.: Comments-oriented blog summarization by sentence extraction. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, pp. 901–904. ACM (2007)
29. SparkNotes: SparkNote on Animal Farm (2007), `http://www.sparknotes.com/lit/animalfarm/` (accessed January 3, 2013)
30. Orwell, G.: Animal Farm (1945), `http://gutenberg.net.au/ebooks01/0100011.txt` (accessed January 3, 2013)
31. Dyke, G., Adamson, D., Howley, I., Rosé, C.: Enhancing scientific reasoning and explanation skills with conversational agents. Submitted to the IEEE Journal on Transactions on Learning Technologies Special Issue on Learning Systems for Science and Technology Education