Pedagogical Evaluation of Automatically Generated Questions

Karen Mazidi and Rodney D. Nielsen

HiLT Lab., University of North Texas, Denton TX
KarenMazidi@my.unt.edu, Rodney.Nielsen@unt.edu

Abstract. Automatic Question Generation from text is a critical component of educational technology applications such as Intelligent Tutoring Systems. We describe an automatic question generator that uses semantic-based templates. We evaluate the system along with two comparable systems for both linguistic quality and pedagogical value of generated questions and find that our system outperforms prior work.

Keywords: question generation, syntactic, semantic, pedagogy.

1 Introduction

This work evaluates three automatic question generation systems which have a common aim: to assist students in remembering and understanding what they have read. Roediger and Pyc [12] describe studies which show that students who are more frequently asked questions retain significantly more than those who are not. Beck et al. [3] demonstrate that reading comprehension can be boosted with questions that are generated automatically. In creating question generation systems for educational technology applications, a crucial design consideration concerns what kinds of questions should be generated. Graesser, Rus and Cai [7] explore many facets of this consideration, including question taxonomies, purpose of questions, and assumptions behind questions. Another consideration is whether the questions should be answerable from the text. This consideration is addressed by Graesser et al. [6] in the context of information sources: whether the answer comes from the text, student knowledge, or other sources.

Question generation approaches are often classified on a syntactic-semantic continuum. In a syntactic approach, the sentence structure is rearranged and altered to turn declarative sentences into questions. Syntactic examples include early work from Wolfe [13] through recent work from Heilman and Smith [8]. Another syntactic approach, the Ceist system [14], manipulates syntax trees, but the rules are stored externally in templates. Syntactic approaches tend to outnumber semantic approaches as seen in the Question Generation Shared Task and Evaluation Challenge 2010 [4] which received only one paragraph-level, semantic entry[11]. Argawal, Shah and Mannem [1] continue the paragraph-level approach using discourse cues to generate questions of types: why, when, give an example, and yes/no. Another recent semantic approach is Lindberg et al. [10]

S. Trausan-Matu et al. (Eds.): ITS 2014, LNCS 8474, pp. 294–299, 2014.

Table 1. Classification of question generation approaches

	Internal Rules	External Rules
Syntactic Constituents	Heilman and Smith	Ceist
Semantic Constituents	Argawal, Shah, Mannem	Lindberg et al.

which used semantic role labeling combined with templates. This latter approach most closely parallels our own; however, our approach is domain-independent, and our system generates answers as well as questions.

Table 1 is provided as an assist in classifying these various approaches. On one axis, approaches are classified according to whether they are manipulating syntactic or semantic constituents of a sentence. On the other axis, they are classified according to whether the rules for this manipulation are internal to the program or kept externally, as in the form of templates. The examples shown are to provide a general frame of reference, not to imply that any one system entirely fits into one category. Most systems cross the boundary lines of Table 1.

2 Approach

The question generation system presented here utilizes semantic role labels and templates. Sentences are processed by SENNA [5], which provides the tokenizing, pos tagging, syntactic constituency parsing and semantic role labeling, using the 2005 Propbank coding scheme [2]. SENNA produces separate semantic arguments for each predicate in the sentence which are matched with appropriate templates. Question generation patterns use the more common semantic roles A0 (proto-agent), A1 (proto-patient), and A2 - A4 (meaning varies by predicate), as well as the ArgM modifiers: directionals, locatives, manner, purpose, cause, discourse, adverbials, and temporal. Templates contain five fields: (1) the question type identifier, (2) required fields, (3) question frame, (4) answer, and (5) filter fields. Generated questions are stored by the question type identifier for later retrieval by question type/depth. The system at the time of this evaluation had 42 question types. Required fields specify what semantic argument should be present, or absent, and any required verb forms. The answer field specifies which semantic argument is the answer to the question. Filter fields will cause a question to not be generated for conditions such as arguments that do not contain nouns. Filters help prevent generating vague or confusing questions.

Table 3 provides examples of questions and the patterns from which they were generated. The question generated in Example 1 uses the form of the verb found in the source sentence. This template requires arguments A0, A1 and ArgM-locative; fields A1 and ArgM-locative are placed in the question, argument A0 is the answer. The full template also had a required field indicating that the verb must be a form of *be* and a filter that excluded predicate-argument sets that included an A2 argument.

Table 2. Examples of Generated Questions

Example 1. Question Frame: What |verb| |A1| |AM-LOC|? **Source text:** Ice wedging is the main form of mechanical weathering in any climate that regularly cycles above and below the freezing point. **Question:** What is the main form of mechanical weathering in any climate that regularly cycles above and below the freezing point? **Answer:** ice wedging
Example 2. Question Frame: How |do A0| |V| |A1|? **Source text:** By examining the arrangement of these dark absorption lines, astronomers can determine the composition of elements that make up a distant star. **Question:** How do astronomers determine the composition of elements that make up a distant star? **Answer:** by examining the arrangement of these dark absoption lines
Example 3. Question Frame: What happens |if|? **Source text:** If the atoms are pulled apart, potential energy goes up because you are separating particles that attract each other. **Question:** What happens if the atoms are pulled apart? **Answer:** potential energy goes up

In Example 2, there are three fields in the question frame that must be replaced with source sentence text. The |do| field will be replaced by *do*, *did* or *does*, depending on the plurality of the nouns and the tense of the verb. The verb will be in its lexical form. Filters in the full template specify that A0 cannot start with a preposition and A1 cannot start with a personal pronoun. The first filter helps with question naturalness and the latter filter helps avoid vague questions. A required field specifies that the ArgM-manner argument which forms the answer must contain a gerund.

In Example 3, the |if| of the question frame will be replaced with the text from the ArgM-adverbial. The full template specification has a filter which indicates that the ArgM-adverbial must contain nouns. This is another filter for vague questions.

3 Linguistic and Pedagogical Evaluations

For these evaluations, we utilized Amazon's Mechanical Turk service. Previous work by Heilman and Smith [9] demonstrates that satisfactory results can be achieved by submitting work in small batches, and closely monitoring each batch. For these evaluations we set up two separate tasks: a linguistic evaluation and a pedagogical evaluation. For the linguistic evaluation, each worker was asked to read the source sentence and question, then rate the question on a 1 to 3 scale for grammaticality and clarity. For the pedagogical evaluation, workers were asked to consider whether this question would help them remember or understand the meaning of the sentence. For all tasks we requested two workers and submitted the questions in batches of 50 or fewer questions.

For these two evaluation tasks we compiled two corpora representing the domains of social studies and science. The social studies text was taken from SparkNotes *Other Topics*. Five files were randomly chosen representing the following domains: Economics: the money supply, History: American History, Government: Federalism, Philosophy: an overview of John Locke's work, and Civics: the development of the nation-state. These files range in length from 27 to 39 sentences, with an average of 33 sentences. The science text was extracted from middle-school and high-school science textbooks downloaded from ck12.org, a non-profit that creates and freely distributes K-12 STEM material. The files represent the following science domains: Life Science: the body, Chemistry: bonds, Biology: the cell, Physics: matter and energy, and Earth Science: weathering. The science files ranged in length from 53 to 69 sentences, with an average of 60 sentences.

Table 3. Inter-rater agreement for Mechanical Turk workers

	Social Studies		Science	
	Linguistics	Pedagogy	Linguistics	Pedagogy
Mean agreement	0.72	064	0.69	0.62
Pearson's r	0.58	0.46	0.57	0.45

Table 3 shows the inter-rater agreement between two sets of workers over all annotations. Mean agreement is calculated as shown below, where i ranges over the N questions rated by the annotators, $r_{1,i}$ is annotator 1's normalized rating $(rating-1)/2$ for the *ith* question (normalized ratings fall in the range [0,1]). We also provide Pearson's correlation coefficient numbers, which indicate a strong positive relationship[1] and are statistically significant, $p < 0.001$.

$$1 - \frac{1}{N}\sum_{i=1}^{N} |r_{1,i} - r_{2,i}| \tag{1}$$

The evaluations described here compare the questions generated by the system described in this paper (M&N), Heilman and Smith's system (H&S), and the Lindberg et al. system (LPN&W). Heilman and Smith's system is available online[2]; David Lindberg graciously shared his code with us. For the following evaluations, 50 questions were randomly selected from all questions generated by each system for a given input file. Table 4 shows the number of questions remaining after a given evaluation filtered out lower-quality questions. The table shows this data for both the social studies and science corpora. For both the linguistic and pedagogical evaluations, the questions that remained were those that received a 3 from one worker, and at least a 2 from the other.

From Table 4, the linguistics evaluation for both data sets are remarkably similar. The average number of questions that remained after applying the linguistics filter to the social studies data was 28, 30, 37 (H&S, LPN&W, M&N),

[1] http://faculty.quinnipiac.edu/libarts/polsci/statistics.html
[2] http://www.ark.cs.cmu.edu/mheilman/questions/

Table 4. Number of acceptable questions for social studies and science corpora

Social Studies		Linguistic Evaluation			Pedagogical Evaluation		
File	Questions	H&S	LPN&W	M&N	H&S	LPN&W	M&N
money	50	36	42	45	18	22	22
amhist	50	29	37	36	22	15	20
federalism	50	27	28	40	11	5	20
locke	50	19	21	30	6	8	10
state	50	27	21	32	11	10	13
Average	50	27.6	29.8	36.6	13.6	12	17
Percent		55.2	59.6	**73.2**	27.2	24.0	**34.0**
Science		Linguistic Evaluation			Pedagogical Evaluation		
File	Questions	H&S	LPN&W	M&N	H&S	LPN&W	M&N
body	50	33	27	42	23	14	25
bonds	50	30	34	31	16	19	19
cell	50	30	26	37	20	14	25
matter	50	25	32	32	12	17	18
weathering	50	18	31	38	9	21	26
Average	50	27.2	30	36	16	17	22.6
Percent		54.4	60.0	**72.0**	32.0	34.0	**45.2**

and for the science data: 27, 30, 36. This speaks both to the consistency of all 3 systems across domains, and to the validity of using MTurk for this evaluation.

Discussion. The question generation systems described in this work begin with expository text. Our system takes this input directly into SENNA. The Heilman and Smith system performs NLP transformations on the input text in order to simplify complex sentences, which they note is "particularly prone to errors"[8]. Using a semantic role labeler essentially performs this simplification itself since it identifies semantic arguments for each predicate in the sentence even within subordinate clauses. The Lindberg et al. system likewise did not perform sentence simplification because they note that important semantic content can be lost, such as temporal information in prepositional phrases [10].

An additional advantage of semantic role labeling is that it can help identify the most salient aspects of a sentence. From: *As the ball gains height, it regains potential energy because of gravity*, a syntactic approach generates the question: What regains potential energy because of gravity as the ball gains height? In contrast, our approach identifies an ArgM-causation argument and can generate a deeper question: Why does the ball regain potential energy?

Heilman and Smith's system provides the answer as well as the generated question, as does our system. The Lindberg et al. system does not provide answers which frees it to ask questions that may not be directly answerable from a sentence. Whether or not this is desirable may depend upon the application.

4 Conclusion

We have evaluated three question generation systems in terms of both the linguistic quality of the produced questions, as well as their pedagogical utility. These types of question generation systems can be integrated into educational technology applications such as Intelligent Tutoring Systems, in order to ensure that students engage deeply with the material. Our system outperformed prior work in both the linguistic and pedagogical evaluations.

Acknowledgements. This research was supported by the Institute of Education Sciences, U.S. Dept. of Ed., Grant R305A120808 to UNT. The opinions expressed are those of the authors.

References

1. Agarwal, M., Shah, R., Mannem, P.: Automatic question generation using discourse cues. In: Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 1–9. Association for Computational Linguistics (2011)
2. Babko-Malaya, O.: Propbank annotation guidelines (2005), `http://www.verbs`
3. Beck, J.E., Mostow, J., Bey, J.: Can automated questions scaffold childrens reading comprehension? In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 478–490. Springer, Heidelberg (2004)
4. Boyer, K., Piwek, P. (eds.): Proc. QG2010: The Third Workshop on Question Generation, Pittsburgh, PA (2010)
5. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Machine Language Research 12, 2493–2537 (2011)
6. Graesser, A.C., Rus, V., Cai, Z., Hu, X.: Question answering and generation. In: McCarthy, P.M., Boonthum, C. (eds.) Applied NLP. IGI Global, Hershey (2011)
7. Graesser, A.C., Rus, V., Cai, Z.: Question classification schemes. In: Proc. WS of the QGSTEC (2008)
8. Heilman, M., Smith, N.A.: Question generation via overgeneration transformations and ranking. No. CMU-LTI-09-013. Carnegie-Mellon University, Pittsburgh (2009)
9. Heilman, M., Smith, N.A.: Rating computer-generated questions with Mechanical Turk. In: Proc. of the NAACL HLT 2010 Workshop CSLDAMT, pp. 35–40. Association for Computational Linguistics (2010)
10. Lindberg, D., Popowich, F., Nesbit, J., Winne, P.: Generating natural language questions to support learning on-line. In: Proc. 14th European Workshop NLG, pp. 105–114. Association for Computational Linguistics (2013)
11. Mannem, P., Prasad, R., Joshi, A.: Question generation from paragraphs at UPenn. In: Proc. QG2010: The Third Workshop on Question Generation, Pittsburgh, PA, pp. 84–91 (2010)
12. Roediger, H.L., Pyc, M.A.: Inexpensive techniques to improve education. J. Applied Research in Memory and Cognition 1(4), 242–248 (2012)
13. Wolfe, J.H.: Automatic question generation from text. ACM SIGCUE Outlook 10(SI), 104–112 (1976)
14. Wyse, B., Piwek, P.: Generating questions from openlearn study units (2009)