Rating Computer-Generated Questions with Mechanical Turk

Michael Heilman
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
mheilman@cs.cmu.edu

Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
nasmith@cs.cmu.edu

Abstract

We use Amazon Mechanical Turk to rate computer-generated reading comprehension questions about Wikipedia articles. Such application-specific ratings can be used to train statistical rankers to improve systems' final output, or to evaluate technologies that generate natural language. We discuss the question rating scheme we developed, assess the quality of the ratings that we gathered through Amazon Mechanical Turk, and show evidence that these ratings can be used to improve question generation.

1 Introduction

This paper discusses the use of Amazon Mechanical Turk (MTurk) to rate computer-generated reading comprehension questions about Wikipedia articles.

We have developed a question generation system (Heilman and Smith, 2009; Heilman and Smith, 2010) that uses the overgenerate-and-rank paradigm (Langkilde and Knight, 1998). In the the overgenerate-and-rank approach, many system-generated outputs are ranked in order to select higher quality outputs. While the approach has had considerable success in natural language generation (Langkilde and Knight, 1998; Walker et al., 2001), it often requires human labels on system output for the purpose of learning to rank. We employ MTurk to reduce the time and cost of acquiring these labels.

For many problems, large labeled datasets do not exist. One alternative is to build rule-based systems, but it is often difficult and time-consuming to accurately encode relevant linguistic knowledge in rules. Another alternative, unsupervised or semi-supervised learning, usually requires clever formulations of bias that guide the learning process (Carroll and Charniak, 1992; Yarowsky, 1995); such intuitions are not always available. Thus, small, application-specific labeled datasets, which can be cheaply constructed using MTurk, may provide considerable benefits by enabling the use of supervised learning.

In addition to using MTurk ratings to train a learned ranking component, we could also use MTurk ratings to evaluate the final top-ranked output of our system. More generally, MTurk can be a useful evaluation tool for systems that output natural language (e.g., systems for natural language generation, summarization, translation). For example, Callison-Burch (2009) used MTurk to evaluate machine translations. MTurk facilitates the efficient measurement and understanding of errors made by such technologies, and could be used to complement automatic evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004).

It is true that, for our task, MTurk workers annotate computer-generated rather than human-generated natural language. Thus, the data will not be as generally useful as other types of annotations, such as parse trees, which could be used to build general purpose syntactic parsers. However, for the reasons described above, we believe the use of MTurk to rate computer-generated output can be useful for the training, development, and evaluation of language technologies.

The remainder of the paper is organized as follows: §2 and §3 briefly describe the question generation system and corpora used in our experiments. §4 provides the details of our rating scheme. §5 discusses the quantity, cost, speed, and quality of the ratings we gathered. §6 presents preliminary experiments showing that the MTurk ratings improve question ranking. Finally, in §7, we conclude.

2 Question Generation System

We use MTurk to improve and evaluate a system for automatic question generation (QG). In our QG approach, hand-crafted rules transform declarative sentences from an input text into a large set of questions (i.e., hundreds per page). This rule system is complemented by a statistical ranker, which ranks questions according to their quality. Currently, we focus on basic linguistic issues and the goal of producing *acceptable* questions—that is, questions that are grammatical, make sense, and are not vague. We believe an educator could select and revise output from the system in order to produce a final set of high-quality, challenging questions.

Our system is described by Heilman and Smith (2010). In that work, we employed a different scheme involving binary judgments of question quality according to various factors such as grammaticality, vagueness, and others. We also employed university students as novice annotators. For the training dataset, only one human rated each question. See Heilman and Smith (2009) for more details.[1]

3 Corpora

In our experiments, we generated questions from 60 articles sampled from the "featured" articles in the English Wikipedia[2] that have between 250 and 2,000 word tokens. This collection provides expository texts written at an adult reading level from a variety of domains, which roughly approximates the prose that a secondary or post-secondary level student would encounter. By choosing from the featured articles, we intended to select well-edited articles about topics of general interest. We then randomly selected 20 questions from each of 60 articles for labeling with MTurk.[3]

[1]We also generated some questions using a technique that replaces pronouns and underspecified noun phrases with antecedent mentions identified by a coreference resolver. We will not provide details about this component here because they are not relevant to our use of MTurk to rate questions. A forthcoming paper will describe these additions.

[2]The English Wikipedia data were downloaded on December 16, 2008 from `http://en.wikipedia.org`

[3]Five questions were later eliminated from this set due to minor implementation changes, the details of which are uninteresting. The final set contained 1,195 questions.

	Rating	Details
1	Bad	The question has major problems.
2	Unacceptable	The question definitely has a minor problem.
3	Borderline	The question might have a problem, but I'm not sure.
4	Acceptable	The question does not have problems.
5	Good	The question is as good as one that a human teacher might write for a reading quiz.

Table 1: The five-point question rating scale.

4 Rating Scheme

This section describes the rating scheme we developed for evaluating the quality of computer-generated questions on MTurk.

Questions were presented independently as single human intelligence tasks (HITs). At the top of the page, raters were given the instructions shown in Figure 1 along with 7 examples of good and bad questions with their appropriate ratings. Below the instructions and examples was an excerpt from the source text consisting of up to 5 sentences of context, ending with the primary sentence that the question was generated from. The question to be rated then followed.

Below each question was the five-point rating scale shown in Table 1. Workers were required to select a single rating by clicking a radio button. At the bottom of the page, the entire source article text was given, in case the worker felt it was necessary to refer back to more context.

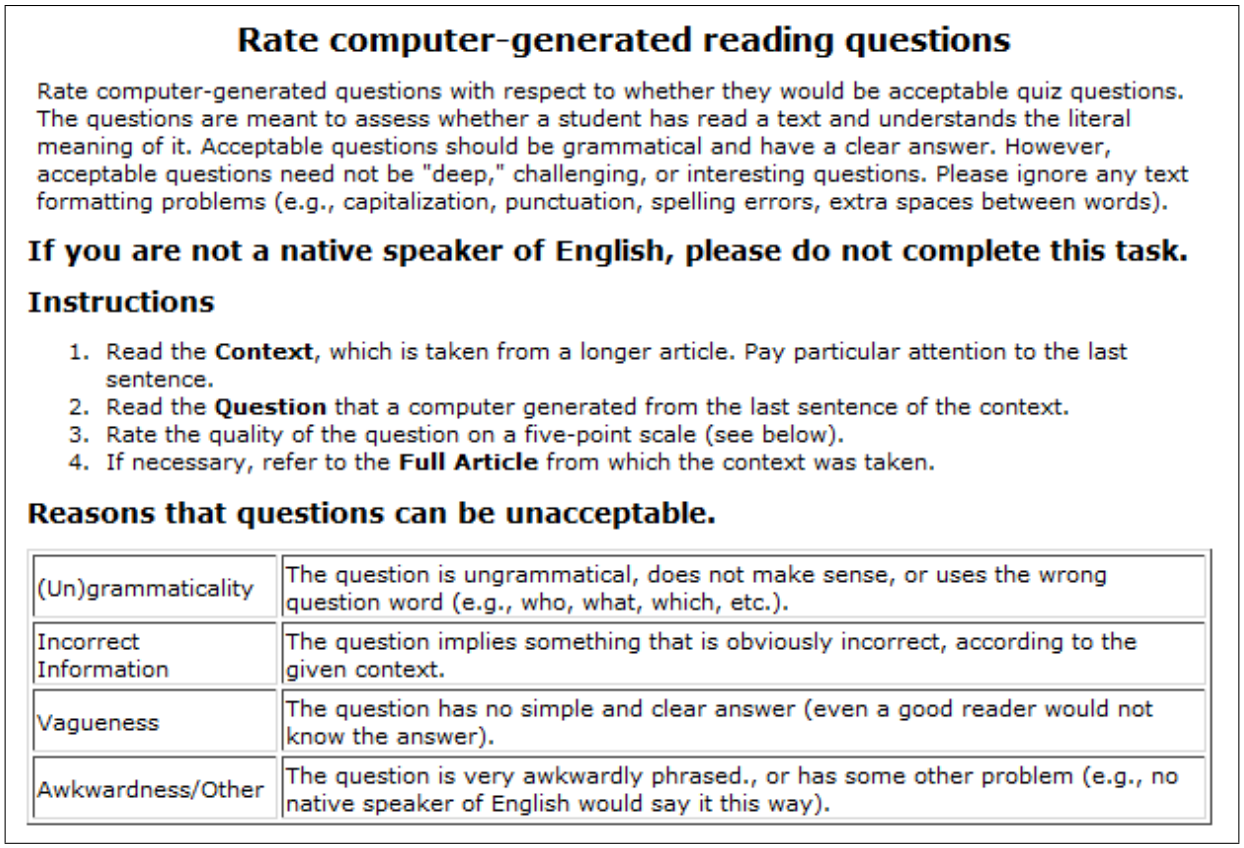
We paid 5 cents per rating,[4] and each question was rated by five workers. With the 10% commission charge by Amazon, each question cost 27.5 cents.

The final rating value was computed by taking the arithmetic mean of the ratings. Table 2 provides some examples of questions and their mean ratings.

4.1 Monitoring Turker Ratings

During some pilot tests, we found that it was particularly important to set some qualification criteria for workers. Specifically, we only allowed workers

[4]Given the average time spent per HIT, the pay rate can be extrapolated to $5–10 per hour.



Rate computer-generated reading questions

Rate computer-generated questions with respect to whether they would be acceptable quiz questions. The questions are meant to assess whether a student has read a text and understands the literal meaning of it. Acceptable questions should be grammatical and have a clear answer. However, acceptable questions need not be "deep," challenging, or interesting questions. Please ignore any text formatting problems (e.g., capitalization, punctuation, spelling errors, extra spaces between words).

If you are not a native speaker of English, please do not complete this task.

Instructions

1. Read the **Context**, which is taken from a longer article. Pay particular attention to the last sentence.
2. Read the **Question** that a computer generated from the last sentence of the context.
3. Rate the quality of the question on a five-point scale (see below).
4. If necessary, refer to the **Full Article** from which the context was taken.

Reasons that questions can be unacceptable.

(Un)grammaticality	The question is ungrammatical, does not make sense, or uses the wrong question word (e.g., who, what, which, etc.).
Incorrect Information	The question implies something that is obviously incorrect, according to the given context.
Vagueness	The question has no simple and clear answer (even a good reader would not know the answer).
Awkwardness/Other	The question is very awkwardly phrased., or has some other problem (e.g., no native speaker of English would say it this way).

Figure 1: A screenshot of the instructions given to workers.

who had completed at least 50 previously accepted HITs. We also required that at least 95% of workers' previous submissions had been accepted.

We also submitted HITs in batches of 100 to 500 so that we could more closely monitor the process.

In addition, we performed a limited amount of semi-automated monitoring of the ratings, and rejected work from workers who were clearly randomly clicking on answers or not following the rating scheme properly. We tried to err on the side of accepting bad work. After all ratings for a batch of questions were received, we calculated for each worker the number of ratings submitted, the average time spent on each question, the average rating, and the correlation of the worker's rating with the mean of the other 4 ratings. We used a combination of these statistics to identify extremely bad workers (e.g., ones who had negative correlations with other workers and spent less than 10 seconds per question). If some of the ratings for a question were rejected, then the HIT was "extended" in order to receive 5 ratings.

5 Quantity, Cost, Speed, and Quality

This section discusses the quantity and quality of the question ratings we received from MTurk.

5.1 Quantity and Cost of Ratings

We received 5 ratings each for 1,200 questions, costing a total of $330. 178 workers participated. Workers submitted 33.9 ratings on average (s.d. = 58.0). The distribution of ratings per worker was highly skewed, such that a handful of workers submitted 100 or more ratings (max = 395). The ratings from these who submitted more than 100 ratings seemed to be slightly lower in quality but still acceptable. The median number of ratings per worker was 11.

5.2 Speed of Ratings

Ratings were received very quickly once the HITs were submitted. Figure 2 shows the cumulative number of ratings received for a batch of questions,

Source Text Excerpt	Question	Rating
MD 36 serves as the main road through the Georges Creek Valley, a region which is historically known for coal mining, and has been designated by MDSHA as part of the Coal Heritage Scenic Byway.	*Which part has MD 36 been designated by MDSHA as?*	1.4
He worked further on the story with the Soviet author Isaac Babel, but no material was ever published or released from their collaboration, and the production of Bezhin Meadow came to an end.	*What did the production of Bezhin Meadow come to?*	2.0
The design was lethal, successful and much imitated, and remains one of the definitive weapons of World War II.	*Does the design remain one of the definitive weapons of World War II?*	2.8
Francium was discovered by Marguerite Perey in France (from which the element takes its name) in 1939.	*Where was Francium discovered by Marguerite Perey in 1939?*	3.8
Lazare Ponticelli was the longest-surviving officially recognized veteran...Although he attempted to remain with his French regiment, he eventually enlisted in...	*Did Lazare Ponticelli attempt to remain with his French regiment?*	4.4

Table 2: Example computer-generated questions, along with their mean ratings from Mechanical Turk.

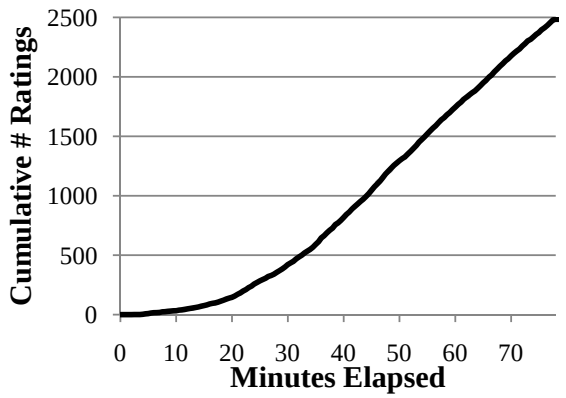


Figure 2: The cumulative number of ratings submitted by MTurk workers over time, for a batch of 497 questions posted simultaneously (there are 5 ratings per question).

indicating that more than 1,000 ratings were received per hour.

5.3 Quality of Ratings

We evaluated inter-rater agreement by having the first author and an independent judge rate a random sample of 40 questions from 4 articles. The independent judge was a computational linguist. The Pearson correlation coefficient between the first author's ratings and the mean ratings from MTurk workers was $r = 0.79$, which is fairly strong though not ideal. The correlation between the independent judge's ratings and the MTurk workers was $r = 0.74$. These fairly strong positive correlations between the MTurk ratings and the two human judges provide evidence that the rating scheme is consistent and well-defined. The results also agree with Snow et al. (2008), who found that aggregating labels from 3 to 7 workers often provides expert levels of agreement. Interestingly, the agreement between the two human raters was somewhat lower ($r = 0.65$), suggesting that aggregated labels from a crowd of MTurk workers can be more reliable than individual humans.[5]

6 Using Labeled Data to Improve Question Ranking

In this section, we provide some preliminary results to demonstrate that MTurk ratings can be used for learning to rank QG output.

First, we briefly characterize the quality of *unranked* output. Figure 3 shows a histogram of the mean MTurk ratings for the 1,195 questions, showing that only a relatively small fraction of the questions created by the overgenerating steps of our system are acceptable: 12.9% when using 3.5 as the threshold for acceptability.

However, ranking can lead to substantially higher levels of quality in the top-ranked questions, which

[5] We also converted the ratings into binary values based on whether they exceeded a threshold of 3.5. After this conversion to a nominal scale, we computed a Cohen's κ of 0.54, which indicates "moderate" agreement (Landis and Koch, 1977).

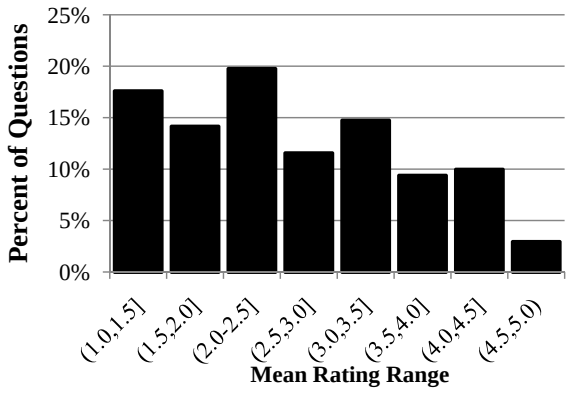


Figure 3: The distribution of the 1,195 question ratings.

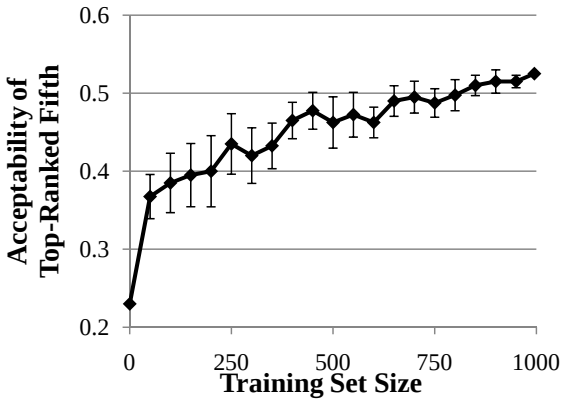


Figure 4: A graph of the acceptability of top-ranked questions when datasets of increasing size are used to train a statistical question ranker. Error bars show 95% confidence intervals computed from the 10 runs of the sampling process.

might be presented first in a user interface. Therefore, we investigated how many MTurk-rated questions are needed to train an effective statistical question ranker. Our ranking model is essentially the same as the one used by Heilman and Smith (2010). Rather than logistic regression, which we used previously, here we use a linear regression with ℓ_2 regularization to account for the ordinal scale of the averaged question ratings. We set the regularization parameter through cross-validation with the training data.

The regression includes all of the features described by Heilman and Smith (2010). It includes features for sentence lengths, whether the question includes various WH words, whether certain syntactic transformations performed during QG, whether negation words are present in questions, how many times various parts of speech appeared, and others. It also includes some additional coreference features for parts of speech and lengths of noun phrase mentions and their antecedents.[6] In all, the ranker includes 326 features.

For our experiments, we set aside a randomly chosen 200 of the 1,195 rated questions as a test set. We then trained statistical rankers on randomly sampled subsets of the remaining questions, from size $N = 50$ up to $N = 995$. For each value of N, we used the ranker trained on that amount of data to rank the 200 test questions. We then computed the percentage of the top fifth of the ranked test set questions with a mean rating above 3.5. For each N less than 995, we repeated the entire sampling, training, and ranking process 10 times and averaged the results. (We used the same 200 question test set throughout the process.)

Figure 4 presents the results, with the acceptability of unranked questions (23%) included at $N = 0$ for comparison. We see that ranking more than doubles the acceptability of the top-ranked questions, consistent with findings from Heilman and Smith (2010). It appears that ranking performance improves as more training data are used. When 650 examples were used, 49% of the top-ranked questions were acceptable. Ranking performance appears to level off somewhat when more than 650 training examples are used. However, we speculate that if the model included more fine-grained features, the value of additional labeled data might increase.[7]

7 Conclusion

In this paper, we used MTurk to gather quality ratings for computer-generated questions. We pre-

[6] Since these additional coreference features are not immediately relevant to this work, we will not describe them fully here. A forthcoming paper will describe them in more detail.

[7] To directly compare the ranker's predictions to the correlations presented in §5.3, we computed a correlation coefficient between the test set ratings from MTurk and the ratings predicted by the ranker when it was trained on all 995 training examples. The coefficient was $r = 0.36$, which is statistically significant ($p < .001$) but suggests that there is substantial room for improvement in the ranking model.

sented a question rating scheme, and found high levels of inter-rater agreement ($r \geq 0.74$) between ratings from reliable humans and ratings from MTurk. We also showed that ratings can be gathered from MTurk quickly (more than 1,000 per hour) and cheaply (less than 30 cents per question).

While ratings of computer-generated language are not as generally useful as, for example, annotations of the syntactic structure of human-generated language, many research paradigms involving the automatic generation of language may be able to benefit from using MTurk to quickly and cheaply evaluate ongoing work. Also, we demonstrated that such ratings can be used in an overgenerate-and-rank strategy to greatly improve the quality of a system's top-ranked output.

References

C. Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proc. of EMNLP*.

G. Carroll and E. Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. Technical report, Brown University.

M. Heilman and N. A. Smith. 2009. Question generation via overgenerating transformations and ranking. Technical Report CMU-LTI-09-013, Language Technologies Institute, Carnegie Mellon University.

M. Heilman and N. A. Smith. 2010. Good question! statistical ranking for question generation. In *Proc. of NAACL-HLT*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33.

I. Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proc. of ACL*.

C. Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proc. of Workshop on Text Summarization*.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP*.

M. A. Walker, O. Rambow, and M. Rogati. 2001. Spot: a trainable sentence planner. In *Proc. of NAACL*.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of ACL*.