Can Summarization Improve Question Quality?

Bethany Griswold, University of Colorado Colorado Springs

***Abstract*— This project focuses on the generation of questions based on Wikipedia articles, as well as the ranking of the questions based on the content therein. A Wikipedia article was first processed through a question generator, which ranked questions generated based on grammatical correctness and comprehensibility of the question. Named entities throughout the Wikipedia article were counted, consolidated, and ranked based on the number of times the entities occurred in the article. Then each question was ranked based on the number and significance of named entities within the sentence. The score given to a question for content relevance was averaged with the score for grammar and comprehensibility, and then a final ranking was done. Another system was also done where the Page Rank algorithm was used on the questions with presence of named entities used as the feature vector, and then the score given to a sentence for page rank was averaged with the grammar score and ranking was done based on the resulting scores. The resulting set of ranked questions were compared against each other and against the results of the Question Generator itself.**

I. Introduction

Wikipedia, a website devoted to the collection and maintenance of a vast amount of knowledge, has in recent years increased in its reliability as an on-line reference. Many other web-based document sites have also emerged as potential educational resources for teachers. However, unlike traditional textbooks, most of these do not have practice and content questions that can be used to test the knowledge and understanding of students who have read the document [1]. Automatic generation of such questions is an interesting and useful problem. Questions will be formulated using a pre-existing system by Heilman [1]. This is how the algorithm is described:

"Each of the sentences from the source text is expanded into a set of derived declarative sentences (which also includes the original sentence) by altering lexical items, syntactic structure, and semantics. [...] a set of transformations derive a simpler form of the source sentence by removing phrase types such as leading conjunctions, sentence-level modifying phrases, and appositives. [...] [The] implementation also extracts a set of declarative sentences from any finite clauses, relative clauses, appositives, and participial phrases that appear in the source sentence. [...] In the second step, the declarative sentences derived in step 1 are transformed into sets of questions by a sequence of well-defined syntactic and lexical transformations (subject-auxiliary inversion, WH -movement, etc.). It identifies the answer phrases which may be targets for WH-movement and converts them into question phrases. [...] The transformation from answer to question is achieved by applying a series of general-purpose rules. [...] Eight Tregex expressions mark phrases that cannot be answer phrases due to WH-movement constraints. [...] We iteratively remove each possible answer phrase and generate possible question phrases from it. [...] Each question is scored according to features of the source sentence, the input sentence, the question, and the transformations used in its generation."

This process is shown in Figure 1.

Ranking of content will be done by calculating significance of named entities and their density within a sentence and averaging that score with the score given to each sentence by Heilman's algorithm. We will also use a Page Rank algorithm with feature vectors of Named Entities within the article as an additional scoring method for relevance. These entities will be taken from data extracted using a pre-existing system which has succeeded in extracting temporal and geospatial information [2].

II. Motivation

Recent methods of automatically generating questions do so by means of direct sentence manipulation to produce natural language questions in English [3]. Sentences in a document are manipulated to replace known entities with question words (who, what, when). The sentence is then rearranged by the system so that it is more likely to be grammatically correct, and the most likely to be grammatically correct are displayed to a user to choose from. By contrast, our method attempts first to obtain the sentences most likely to be dense with content most central to the article, and focuses on ranking by content importance and relevance. The question generator will still rank for grammatical correctness, but then the questions will be re-ranked to account for relevance.

III. Prior Work

There has been a variety of work previously on question generation from textual content in the past. One method is based on manipulating the structure of each sentence in an article to generate one or more questions, then ranking the questions based on grammatical correctness and importance. This is the "over-generate and rank" method, on which this research will be partially based [3]. Another method is question generation for the particular domain of vocabulary assessment [4]. There are also some tools and research on which this project relies. The first is the Geografikos system, which extracts temporal and geospatial data and then tags sentences to indicate when and where they happened [2] [5]. We are also following prior work of using summarization for feature selection as a first step to a further goal. Where our goal is question generation sorted by importance, it has previously been used for text categorization [6]. We also used the Page Rank algorithm as an additional content importance ranking system [7], [8].

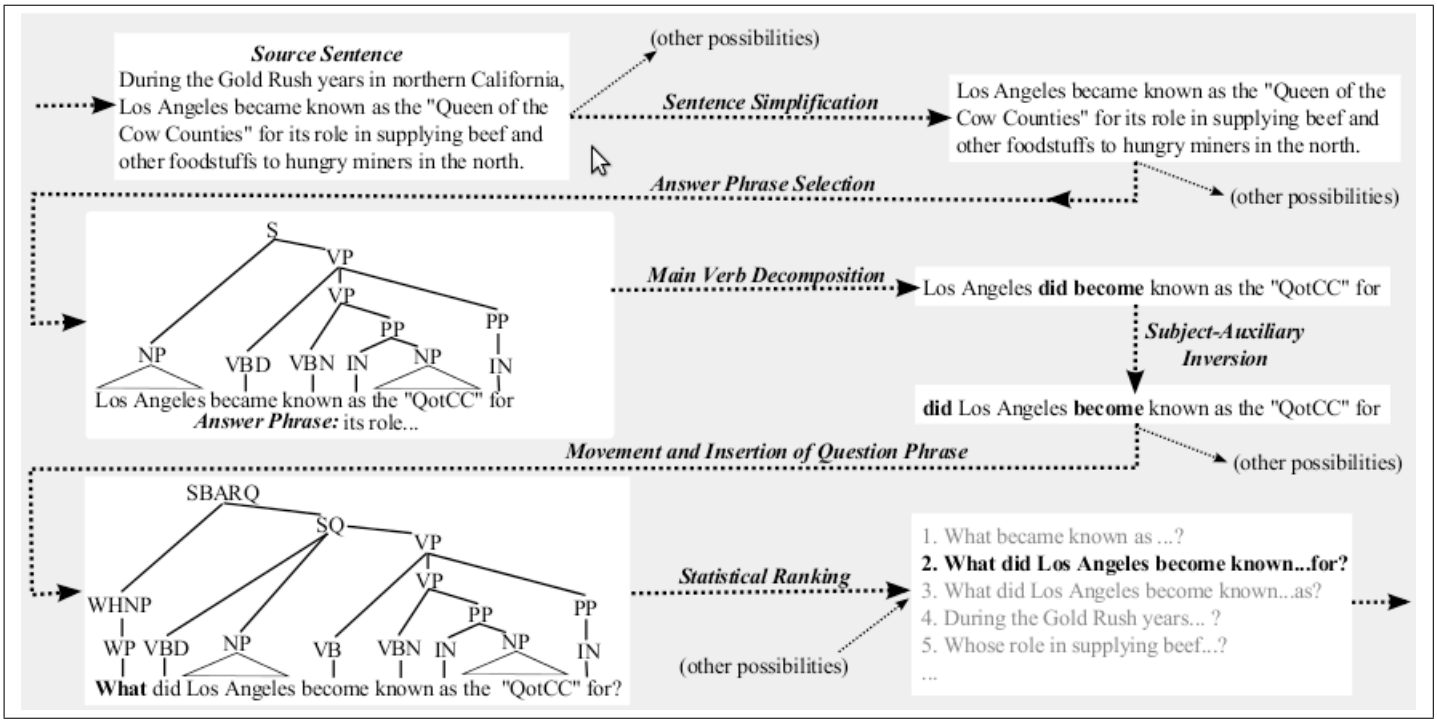

Fig. 1. Heilman's Overgenerate and Rank Process

IV. Project Goals

Our ultimate goal is to generate concise questions about historical articles which are ranked based on the strength of their relation to the core ideas of an article.

At the start, we strip a Wikipedia article down to plain text. As an example, use an article on the Battle of Fredericksburg. Here is an excerpt from that article:

"*The battle was the result of an effort by the Union Army to regain the initiative in its struggle against Lee's smaller but more aggressive army. Burnside was appointed commander of the Army of the Potomac in November, replacing Maj. Gen. George B. McClellan. Although McClellan had stopped Lee at the Battle of Antietam in September, President of the United States—President Abraham Lincoln believed he lacked decisiveness, did not pursue and destroy Lee's army in Maryland, and wasted excessive time reorganizing and re-equipping his army following major battles.*"

The question generation tool created by Heilman generates questions and ranks them based on the structure of the question. Following are the top four ranked questions from Heilman's question generator, given the article on the Battle of Fredericksburg:

- When was Burnside appointed commander of the Army of the Potomac?
- Who was Burnside appointed in November?
- What was Burnside appointed commander of of the Potomac in November?
- What was Burnside appointed commander of the Army in November?

All of these questions are generated from the same sentence, but only the first one is really a grammatically good question. None of them are questions that address a very central concept of what, when, and where the battle was or why it happened. From a summarizer, we would hope to preserve as important a sentence such as the first in the given paragraph:

"*The battle was the result of an effort by the Union Army to regain the initiative in its struggle against Lee's smaller but more aggressive army.*"

Such a question could be, "*What was the battle a result of?*"

We have multiple potential methods to try and accomplish this. We used two methods of ranking for relevance based on named entities. For each of these, the article is process through Heilman's question generator first, in order to get the grammar score. The first analyzed the text for named entities, ranked them, and used their presence within a given sentence to determine the importance of that sentence to the text overall. This is done by counting up the number of times a named entity is mentioned within an article and giving that named entity a score of that number. Then each sentence is given a score determined by all the scores of the named entities within that sentence divided by the length of the sentence overall. This score is averaged with the grammar score determined by Heilman's algorithm and the resulting questions are ranked, with the highest scored question first and the lowest scored question last. The second method uses the Page Rank algorithm, which determines relevant and similar sentences based on the similarity of what named entities each has. These processes are shown in Figure 2. The diagrams show that following each process, the score for relevance is averaged with Heilman's score. We also used a fairly simple method for comparison, which consists of processing the entire text through a summarizer and then processing the summary through a question generator, as in Figure 3.

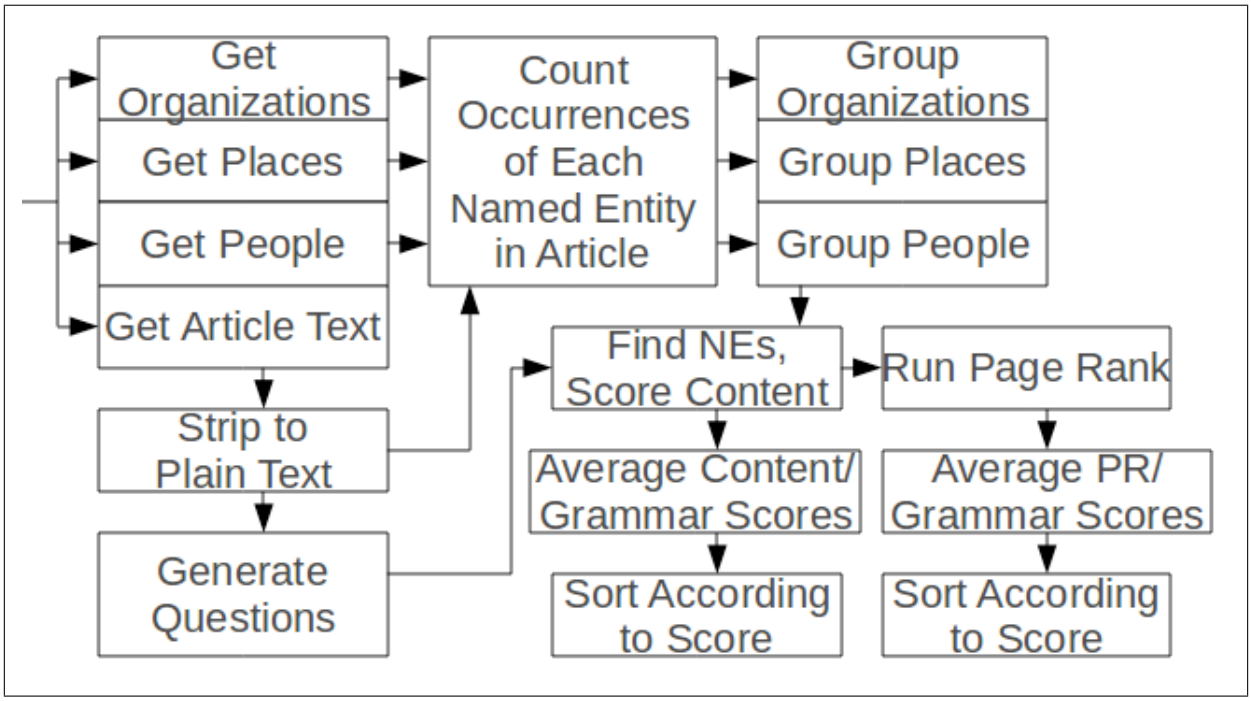

Fig. 2. Ranking Process for NE Based Content Ranking and Page Rank Based Ranking. Any summarization algorithm can be run on a body of text before sending through this process in attempt to acheive better results.

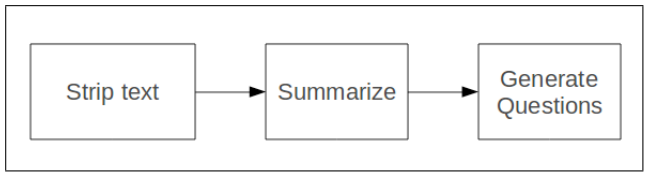

Fig. 3. Summary Based Ranking Process

V. Evaluation

The top 50, 25, and 10 questions, fair numbers of questions for a quiz or test, produced by our six current methods of ranking were rated by one person each on four different features. The first feature is difficulty, which we are testing to see if there is a trend between method of question generation and difficulty of questions. We hope we can pick out questions out of a corpus for a particular grade level. The second feature is relevance, which tests whether or not understanding of the particular question is important or vital to understanding of the article as a whole. We hypothesize that most questions rated highly for content by our algorithm should on average be rated highly for the relevance feature by our evaluators. The third feature is precision, which is the degree to which the evaluator understands what the question is asking and feels that the question is comprehensible and grammatically. We feel that high ratings for precision will closely correspond to a high ranking in Heilman's algorithm.

All evaluators were given the full text of the article on the Battle of Fredericksburg, along with the top 50 questions produced by one of the methods of ranking. Each ranking method was given to one person only. We decided not to allow one person to evaluate more than one set of questions, as we were concerned that a previous set of questions they had read would influence their opinion of the current set of questions. We also did not tell them the ranking method we used, in case assumption of one set having a better ranking method would influence their evaluation of the questions.

There is a bit of a conflict in doing these evaluations, as strong readers may be best suited to assess vagueness, relevance, and completeness, but their high reading level may cause their judgments of difficulty to be very low. We do not have the resources, however, to poll a wide variety of reading levels, so while this point must still be evaluated, the bias of the evaluators must be taken into account. We may also be able to get better results if we are able to have each question set evaluated by a greater number of people, so that the ratings are less influenced by individual skill level.

The methods of question ranking which we decided to give out for evaluation are as follows:

- Question generation from the full text and no further ranking.
- Question generation from the first and last paragraph summary of the full text and no further ranking.
- Question generation from the full text, followed by Named Entity based content ranking.
- Question generation from the first and last paragraph summary of the full text, followed by Named Entity based ranking.
- Question generation from the full text, followed by Page Rank based content ranking.
- Question generation from the first and last paragraph summary of the full text, followed by Page Rank based content ranking.

Different Summarizers could be used in place of the very simple reduction to the first and last paragraph of the text, if a more intelligent summarization system was desired.

VI. Results

One observation given to us by some of the evaluators was that questions generated that required simply yes or no answers were almost too easy to be good questions. Since this is not accounted for in the rankings, a binary of whether something is an acceptable or good question, we may want to do some evaluations in the future that do address this, or otherwise use the feature of Heilman's algorithm that limits output to questions that cannot be answered by true or false before doing any more ranking and evaluation.

The average scores on each quality for the top 50, 25, and 10 generated questions are shown in the graph in Figure 4. Set 1 is the question generation from the full text followed by Named Entity ranking. Set 2 is the set produced by summarizing the text by using the first and last paragraph before generating questions from the summary and then processing the questions through NE based raking. Set 3 was obtained by processing the full text through the question generator, but no other ranking was done. Set 4 is similar, but the text undergoes first and last paragraph summarization first. Set 5 is the full text processed through the question generator and then ranked using the Page Rank algorithm. Set 6 also uses the Page Rank algorithm, but starts out with a first and last paragraph summary.

There were some interesting features of the rankings in comparison to each other. We anticipated that when summarizing and ranking based on named entities, overall relevance of the top 50 ranked questions to the central points of the article would be improved from algorithms that focused primarily on comprehensible grammar. However, as you can see in Figure 4, this is not always the case. The questions in set 3 were ranked most highly for relevance to the article. However, set 3 was the set produced by Heilman's algorithm only, and nothing was actively done to promote more significant questions in the ranks. In fact, it seems that on average, no feature of the top questions was significantly improved from the base question generation by counting named entities and altering the ranking of questions based on density and significance of named entities or by summarizing the text before question generation. However, compared to a simple summarization before question generation, summarization followed by question generation and NE based content ranking shows a marked improvement in precision, relevance, and completeness, as you can see from set 2 and set 4 in Figure 4.

Seeing that the overall average rating of features did not improve with our added content ranking systems, we also did some analysis of the evaluations on a case-by-case basis, with regards to the trends of how well certain features were ranked as an evaluator goes down the list of questions. We expected that the questions near the top of the list would be higher rated in all features except perhaps difficulty than those near the bottom of the list. However, as you can see in Figures 5 and 6, this is not the case in any generalizable way. Set 3 has only a slight tendency toward both better precision ratings and better

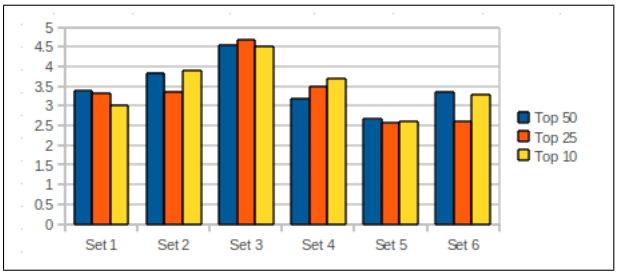

Fig. 5. Relevance scores for all sets of questions generated for the top 50, 25, and 10 questions. Only the questions in set 4 shows improvement of question relevance ratings with increased selectivity of questions.

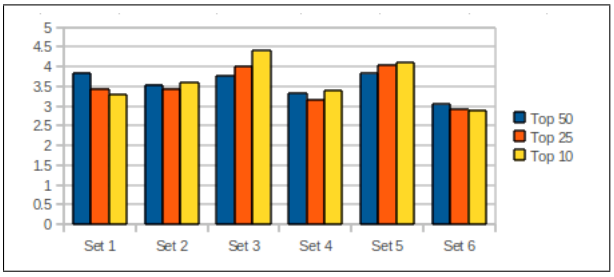

Fig. 6. Precision scores for all sets of questions generated for the top 50, 25, and 10 questions. Sets 3 and 5 are the only sets of questions that show improvement of question precision ratings with increased selectivity of questions.

relevance ratings in earlier questions, and this is, once again, the set that we did not perform any additional rankings on. The failure of all the additionally content ranked sets we believed was likely due to content scores being averaged with grammar scores, so that many of the questions may be very strong in one sense, which compensates enough statistically for weakness in another feature to allow it to be placed high in the rankings. To test this hypothesis, we superimposed the rating for precision, which should be most closely tied to ratings for grammar and comprehensibility, on top of the ratings for relevance, which should be closely tied to algorithmic ranking of content. This is shown in Figure 7. If it was the case that our algorithmic content ranking matched up with human evaluation of content importance or centrality, Heliman's algorithmic grammar and comprehensibility ranking matched up with human evaluation of precision, and the averaging of these features was causing promotion of bad questions, we would see a nearly polar opposite ranking on the other feature for questions rated very high in a single feature late in the rankings or very low on a single feature earlier in the rankings. We once again do not see this in a definite enough trend to make a conclusive statement about whether this is the case. However, it is once again possible that given a high degree of variability in ratings due to variability in skill levels, more evaluators could prove useful in giving a more accurate assessment of each corpus of questions.

Another feature we wanted to consider looking into was whether we could use certain summarization methods to automatically extract questions with a certain level of difficulty from a large corpus of questions. On this task, we got fairly reasonable results. The questions with the highest difficulty came from set 1, which generated questions from the full text

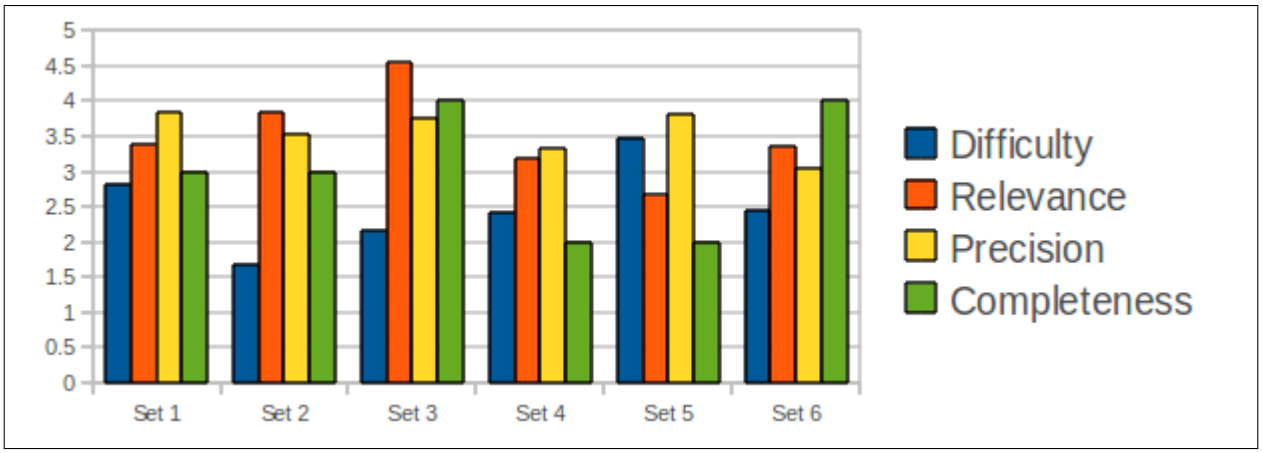

Fig. 4. Average ratings for features of top 50 questions generated by rating methods

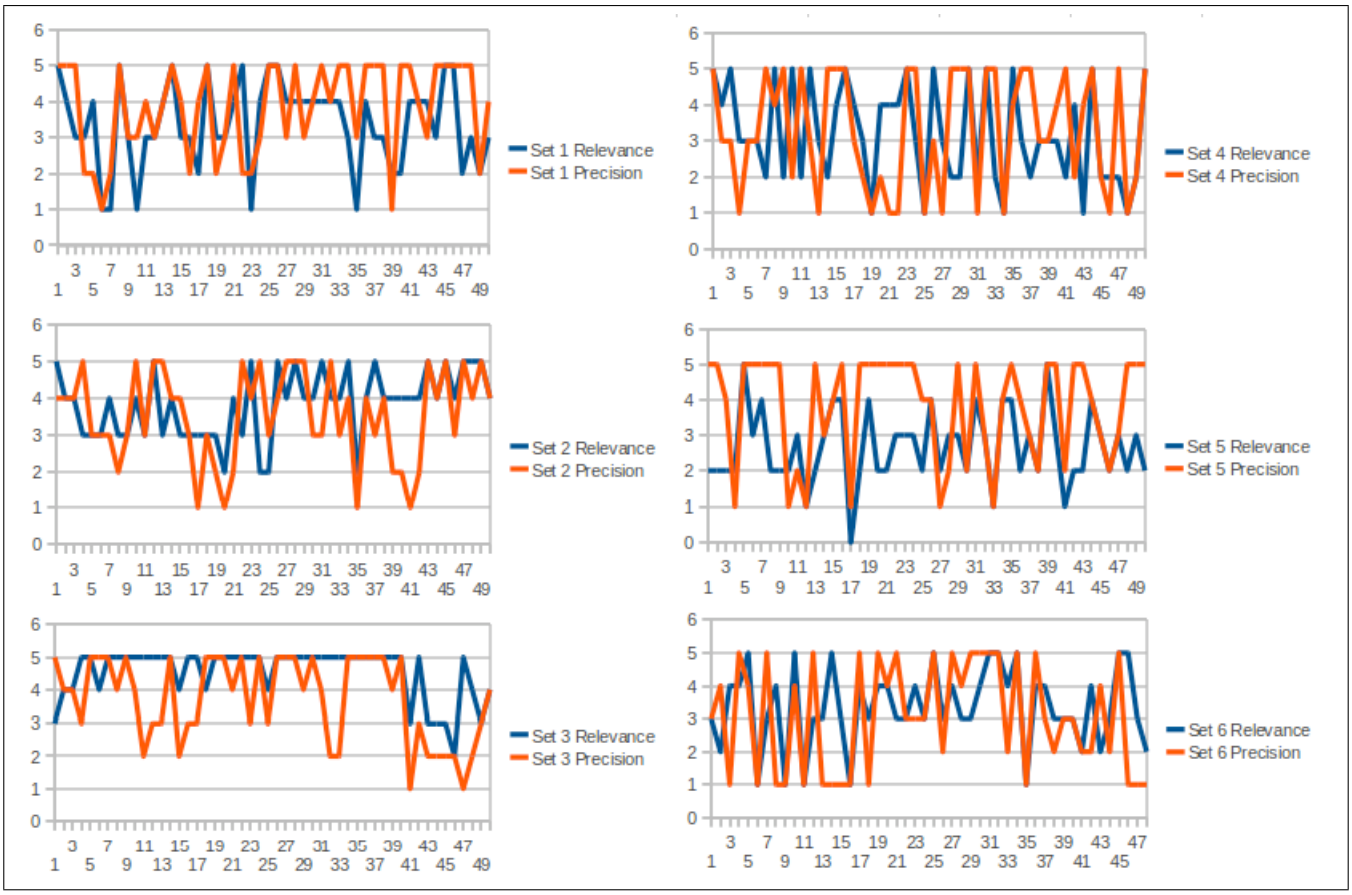

Fig. 7. Precision scores superimposed on top of relevance scores. We supposed that we could be getting low ratings on questions ranked in the top 50 due to a polarity between the grammatical correctness and article relevance giving a question a decent score and rating. While some cases show that low precision scores are counterbalanced by high relevance scores and vice versa, this is by no means an overarching theme in our current data and we cannot assume this is the cause of a lack of downward trend in acceptability of questions in lower rankings of questions.

and then re-ranked based on NE counts. The questions with the lowest difficulty came from set 2, which underwent the same process as set 1, but first was summarized to content only in the first and last paragraph. Set 1 likely has a high degree of difficulty in part because questions were generated from the entire body of the text, where there are many minute details that could be difficult to retain throughout the course of reading the document. Yet this may not be entirely true, since the questions from Set 3 were deemed slightly less difficult than those from set 4, despite that Set 4 was summarized while Set 3 was not and the process for generating and ranking the questions was otherwise identical.

VII. Conclusion

Neither simple summarization before nor content ranking based on Named Entity counts after question generation shows any obvious improvement in question quality, article relevance, or completeness of a question set in evaluating understanding of an article. Only a very slight improvement in precision was seen from the sets with named entity counting and no summarization and page rank ranking and no summarization. This could be due to the averaging of scores for content and grammar quality resulting in promotion of questions that are very good in one feature and very poor in another, but analysis of the ranking data does not give conclusive support for this hypothesis. It is also possible that more sophisticated summarizers could show an improvement, or that using a voting method of rating rather than an average content and grammar score method could eliminate promotion of questions with poor scores in one feature or the other. Further evaluation may also need to be done on our current set of data in order to determine how heavily individual ability played a part in initial rankings of questions, and we may also want to have a binary evaluation of questions being acceptable or unacceptable. Until further evaluation proves otherwise, we must conclude that Named Entity significance and density evaluation on questions does not help to improve the content quality and overall acceptability of generated questions.

VIII. Future Work

To optimize on the different strengths and weaknesses of each method, we hope to combine various methods in a voting system in order to rank key questions highly and attempting to maintain good, comprehensible grammatical structure. It is possible that using a voting system between the top ranked questions using Heilman's algorithm and those ranked by any number of ranking systems based on relevance will be more effective than averaging scores in bringing questions to the top which are both fairly relevant and fairly grammatical. It could avoid retrieving questions that are highly grammatical but completely irrelevant or highly relevant but incomprehensible. Currently, we have a voting algorithm implemented, but several result sets must still be generated from it and evaluations must be gathered from a group of people.

It would also be beneficial to attempt to improve results by using a wider variety of summarizers before question generation. One such summarizer is the MEAD summarizer [9]. Summarization algorithms are meant to compress a document down to the essential and central information, and so more sophisticated summarization than first and last paragraph summarization could potentially result in a drastic improvement in result relevance. However, it's also possible that summarization could cause an increase in overall ambiguity of the questions. This is partially due to there being simply fewer sentences which Heilman's algorithm can rank for grammar and comprehensibility, and partially because summarization inherently removes information, and lack of enough information is what leads to ambiguity.

It is possible that the quality, utility, and relevance of the results and ratings we obtained from our evaluators was substantially decreased by the low number of evaluators we had altogether. In order to gain better results for analysis and ensure that the ability of the individuals rating sets produced by various methods is not disproportionately skewing the data to make it appear that certain methods are better than others, a few things need to be done. The first is that it is essential to have a greater number of people do further ratings on our current results, so that we can do a more accurate evaluation on the usefulness of various methods for articles like that on the Battle of Fredericksburg. Then, we must produce evaluation forms for other articles, which should not be a difficult task, since we have all algorithms in place, and also distribute those to a number of people. It could be that while the basic grammar-only ranking is best for very complex articles like that on the Battle of Fredericksburg for producing acceptable questions, articles that are shorter or longer, simpler or more complex, or more or less technical would have significantly better questions resulting from a different ranking process. If this is the case, these features could be analyzed in the documents and incorporated into the voting algorithm.

Finally, we should attempt to determine what features in a sentence lead to a high level of difficulty. Since the purpose of generating questions is to assist in developing content for education, it would be helpful if it were possible to develop an algorithm that would attempt to sift through a set of questions to find those appropriate for a particular grade level. This may require deeper analysis of not only grammatical structure and Named Entity significance, but also analysis of vocabulary to determine if the general language of a question, or perhaps a full article, is appropriate for a certain age, grade level or reading level.

Acknowledgement

The research reported in this document has been funded partially by NSF grants CNS-0958576 and CNS-0851783.

References

[1] M. Heilman, "Automatic factual question generation from text," Ph.D. dissertation, Carnegie Mellon University, 2011.

[2] R. Chasin, D. Woodward, and J. Kalita, *Machine Intelligence*. Narosa Publishing, 2011, ch. Extracting and Displaying Temporal Entities from Historical Articles, pp. 1–13.

[3] M. Heilman and N. Smith, "Good question! statistical ranking for question generation." Los Angeles: Citeseer, 2010, pp. 609–617.

[4] J. Brown, G. Frishkoff, and M. Eskenazi, "Automatic question generation for vocabulary assessment," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 819–826.
[5] J. Witmer and J. Kalita, "Extracting geospatial entities from wikipedia," in *Third IEEE on Semantic Computing*. Berkely, CA: ICSC 2009, September 2009, pp. 450–457.
[6] A. Kolcz, V. Prabakarmurthi, and J. Kalita, "Summarization as feature selection for text categorization," in *Proceedings of the tenth international conference on Information and knowledge management*. ACM, 2001, pp. 365–370.
[7] G. Erkan and D. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, no. 1, pp. 457–479, 2004.
[8] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in *Proceedings of EMNLP*, vol. 4. Barcelona: ACL, 2004, pp. 404–411.
[9] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang, "MEAD - a platform for multidocument multilingual text summarization," in *LREC 2004*, Lisbon, Portugal, May 2004.