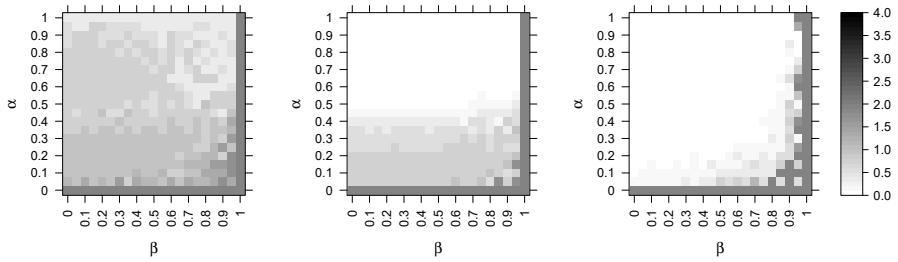

(a) Random baseline (b) Dedicated-quest. baseline (c) Balanced quiz

Method	Mean rank of unread document*	Mean % correct answers
Balanced	1.40 [0.88, 1.96]	49%
Dedic.-qu.	1.60 [1.12, 2.08]	42%
Random	2.00 [1.36, 2.60]	38%

*95% confidence intervals in parentheses.

(d) Human evaluation

Figure 2: (a–c) Simulated evaluation results; colors represent mean rank of unread document (lighter colors are better); definitions of α, β in Sec. 4.2. (d) Human evaluation results. Balanced quizzes are best at detecting documents not read by student.

the answer to q, and that knowing the answer ahead of time (β) is independent of (re)discovering it by reading a document (α).

The probability of answering correctly a question q whose answer appears in n_q documents is thus $p_q = 1 - (1-\alpha)^{n_q}(1-\beta)$, and a document d's expected fraction p_d of questions answered correctly (our ranking criterion) is the average of p_q over the n_d quiz questions that d answers: $p_d = \frac{1}{n_d} \sum_{q \in S:\{d,q\} \in E} p_q$.

The heatmaps in Fig. 2(a–c) plot the mean rank of the skipped document for our quizzes and for the two baselines as a function of α and β. As there are 5 documents, the mean rank can take on values between 0 and 4, with lower values (lighter colors) being better. As expected, the random baseline performs worst. The DQ baseline, on the contrary, reveals the skipped document nearly perfectly when $\alpha \geq 0.5$, but does much more poorly for smaller values of α (i.e., when the simulated student reads sloppily). Our balanced quizzes, finally, show the best performance, with the average rank of the skipped document being 0 for a much wider range of (α, β). Also note that the upper left triangle of the heatmaps is most meaningful, since here the probability of discovering an answer when reading a document is larger than the probability of already knowing the answer before reading any document. In this regime, unlike the two baselines, our quizzes show essentially perfect performance.

Human evaluation. The above simulation study, though useful for initial insights, makes rather strong assumptions. Hence we also conduct an evaluation involving 75 real human quiz takers, recruited via Mechanical Turk. We again use the same 5 immunology-related articles and require $k = 25$ questions per quiz. The 75 quiz takers are split evenly over the three methods. We build one quiz per method, with 25 users taking each quiz. We evaluate each quiz in 5 conditions (5 users per condition): in each condition, a different one of the 5 articles is hidden from the user, the other 4 are shown, thus simulating a scenario where the user skips exactly that document. Users provide their answers as short free-form texts, and we manually determine the correctness of each answer.

Averaging over the 25 instances of each quiz, we obtain the mean rank of the skipped document. Fig. 2(d) shows that it is lowest for our balanced quizzes (1.4), followed by the DQ baseline (1.6), with the random baseline again performing worst (2.0). (Likely due to the small sample size of 25 instances per quiz, the differences are not statistically significant.) The mean ranks are higher than for most (α, β) in our simulations, probably because the quizzes are generally rather hard, with less than half of all questions answered correctly (Fig. 2(d), right column), which induces a lot of noise. Note, however, that our quizzes are easier to answer (49% of questions answered correctly on average) than the baselines (42% and 38%).

5 DISCUSSION AND CONCLUSION

This paper presents a method for combining questions about a document collection into quizzes where questions relate to documents and concepts in a balanced fashion. Our contribution adds to previous work, which has mostly concentrated on generating candidate questions from document collections, whereas our focus is on compiling candidate questions into meaningful quizzes.

Our approach leverages a graph representing the relationships between questions, documents, and concepts, and phrases quiz construction as a node selection problem in this graph. We provide a method for constructing the graph and for selecting a good set of quiz questions using a greedy algorithm. Our results are promising, both in a simulation study and in an evaluation with human quiz takers. Future work should go further both quantitatively and qualitatively, by conducting experiments with more participants as well as with students rather than crowd workers, since the latter might not be representative of real classroom settings.

To further improve our quizzes, future work also should aim to produce sequences, rather than sets, of questions (such that questions build on each other and guide students through the documents in a meaningful order). Finally, it would be useful to extend our algorithm such that it can produce many different quizzes for the same document collection, rather than a single optimal one.

REFERENCES

[1] Q. Guo, C. Kulkarni, A. Kittur, J. Bigham, and E. Brunskill. 2016. Questimator: Generating knowledge assessments for arbitrary topics. In *IJCAI*.
[2] Y. Huang, Y. Tseng, Y. Sun, and M. Chen. 2014. TEDQuiz: Automatic quiz generation for TED talks video clips to assess listening comprehension. In *ICALT*.
[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.
[4] R. Mitkov, L. Ha, and N. Karamanis. 2006. A computer-aided environment for generating multiple-choice test items. *Nat. Lang. Eng.* 12, 2 (2006), 177–194.
[5] G. Nemhauser, L. Wolsey, and M. Fisher. 1978. An analysis of approximations for maximizing submodular set functions – I. *Math. Progr.* 14, 1 (1978), 265–294.
[6] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv:1606.05250* (2016).
[7] F. Rousseau and M. Vazirgiannis. 2015. Main core retention on graph-of-words for single-document keyword extraction. In *ECIR*.
[8] K. Sakaguchi, Y. Arase, and M. Komachi. 2013. Discriminative approach to fill-in-the-blank quiz generation for language learners. In *ACL*.
[9] D. Seyler, M. Yahya, and K. Berberich. 2015. Generating quiz questions from knowledge graphs. In *WWW*.
[10] A. Singh Bhatia, M. Kirti, and S. Saha. 2013. Automatic generation of multiple choice questions using Wikipedia. In *PReMI*.
[11] T. Wang, X. Yuan, and A. Trischler. 2017. A joint model for question answering and question generation. *arXiv:1706.01450* (2017).