When scientists first began experimenting with question-answering problems, there was a lack of datasets available for use to train and test models. To fix this, researchers developed synthetic datasets of triplets, comprised of context, questions, and answers1. However, models created from these datasets typically fail when transitioning to actual data. Hermann et al. addressed this by creating triples from CNN and Daily Mail articles and their summaries; the articles themselves were the context, and multiple questions were created from the summaries by selecting an entity within the summary as the answer, and the remaining text became the question. Since the summaries were not direct copies of the text within the articles, the ability to answer these questions would demonstrate a form of reading comprehension as opposed to recognizing semantic or syntactic similarities. The researchers developed LSTM and transformer models to attempt to answer the questions in their new test dataset, which yielded accuracies ranging from 60% - 70%.

Models developed in this way excel at answering questions with answers that are explicitly stated but fail to perform well on more complex questions2. Welbl et al. sought to address this by creating a set of questions that could only be answered by referencing two or more articles, thereby eliminating questions that could be answered by reading a single sentence3. They did so by creating the “WikiHop” dataset. This is like the news dataset from Hermann et al. in its structure as triples of context, questions, and answers; however, rather than a single article, the context in the WikiHop dataset is comprised of multiple Wikipedia articles. This makes the question-answering problem much harder since the model may not know where to locate the answer, and how to hop from one article to another when searching for the answer. The researchers tested multiple models, but their most effective model was again a form of LSTM neural network. This model had an accuracy of 54.5%. Given that human accuracy on this problem is typically around 85%, there is an opportunity to improve upon this result. Additionally, Clark et al. point out that questions are usually binary predicates and can often be answered without having to “hop” to another article.

There have also been attempts to create models answering standardized test questions similar to the task this paper will attempt. However, past attempts have only been able to use small datasets, leading simpler models to perform better than deeper, more complex models4. For example, Clark et al. attempted a similar task with a training set of only 108 questions (compared to the training set to be used in this paper of size 7,787 questions). The models which performed the best on their test set were basic information retrieval methods or even correlation. While this paper will initially attempt such models as well, it will also include LSTM and transformer networks to attempt to answer the standardized questions.

1. K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt,W. Kay, M. Suleyman, and P. Blunsom.

Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems, pp. 1693–1701, 2015.

2. Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, Oyvind

Tafjord Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge

3. J. Welbl, P. Stenetorp, and S. Riedel. Constructing datasets for multi-hop reading comprehension

across documents. arXiv preprint arXiv:1710.06481, 2017b.

4. P. Clark, O. Etzioni, T. Khot, A. Sabharwal, O. Tafjord, P. D. Turney, and D. Khashabi. Combining

retrieval, statistics, and inference to answer elementary science questions. In AAAI, pp. 2580–2586, 2016.