**Classics’ Comprehension**

Jennifer Philippou

Jennifer.p@berkeley.edu

April 21, 2018

## **Abstract:**

*In this paper we consider a new reading comprehension dataset that incorporates longer bodies of context and stronger answer supervision than prior works. To efficiently handle the larger set of context data with limited computational power and time available, a lightweight text processing pipeline was deployed. With a combination of classic text processing and an LSTM model an accuracy of 16% was achieved. The opportunity to further explore applying cutting edge techniques to this new dataset as well as the opportunity to advance the field of reading comprehension using Lexile scores exists.*

## **1 Introduction**

Teaching computers reading comprehension is a very active area of interest. Access to information on the Internet revolutionized many industries and greatly impacted daily life for most of the population. Recently the technology giants have made great strides in audio comprehension with virtual assistants. The next step in the AI revolution is for computers to better understand large volumes of written text and this will fundamentally alter key areas of information retrieval resulting in cascading effects for the medical, legal, and academic fields.

## **2 Prior Research**

Very few datasets exist to evaluate machine reading comprehension. While a large volume of written works is available, it is difficult to find appropriate questions and corresponding answers to effectively assess the level of understanding. Furthermore, evaluating the complexity of the base corpus is another challenge that is not addressed by many of the datasets that are currently available for research. Table 1 shown below summarizes the features of recent prior question and answer datasets.

Table 1: Summary of Prior Research

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Paper Published** | **Question Size** | **Context**  **Size** | **Context**  **Description** | **Supervision** | **Natural Language Questions** | **Context Complexity Evaluation** |
| MC Test | Oct 2013 | 2,640 | 660 | Paragraph | Y | Y | Y |
| bAbl Children’s Book | April 2016 | 687,343 | 108 Books | Sentence | Y | N | N |
| SQUAD | Oct 2016 | 107,785 | 536 | Paragraph | Y | Y | N |
| MS MACRO | Nov 2016 | 100,000+ | 1M Passages 200k Documents | Paragraphs from several sources | Semi | N | N |
| TriviaQA | May 2017 | 95,000 | 650k (6 per Question) | Multiple Documents | Semi ~4k Questions | Y | N |
| WikiSuggest | Feb 2017 | 3.47M | 3.47M | Long Docs | Semi | N | N |
| Classics Comprehension | NA | 16k+ | 1.7k | Chapters | Y | Y | Y |

(Eunsol, ‘17)(Hill, ’16)(Richardson, ’13)(Rajpurkar, ‘16)(Mandar, ’17)(Nguyen, ‘16)

## **3 Methods**

The Classics’ Comprehension dataset leverages two primary sources: full texts generously made available by the Gutenberg project and questions generated for classroom usage. The questions were downloaded from a platform designed to teach literature, namely the BookRags website. These natural language questions make the dataset distinct from MS MARCO, WikiSuggest, and the Children’s Book test. Each book has one hundred eighty reading comprehension questions and answers. With almost one hundred books considered in the Classics’ Comprehension dataset, it contains over sixteen thousand questions for modeling and that is more than six times the number of questions in the MC Test dataset. In figure 1 you can see a few example questions for

Figure 1: BookRag Question Demo

Example questions:

**1. Where is Alice when she sees the White Rabbit for the first time?**   
On a riverbank   
**2. Alice does NOT think she might end up in what place?**

America

**3. What does Alice think about on her way down the rabbit hole?**

Dinah

Alice’s Adventures in Wonderland (BookRags). These questions are allocated to specific portions of each book; sometimes one chapter and other times several chapters.

## **4 Data Analysis**

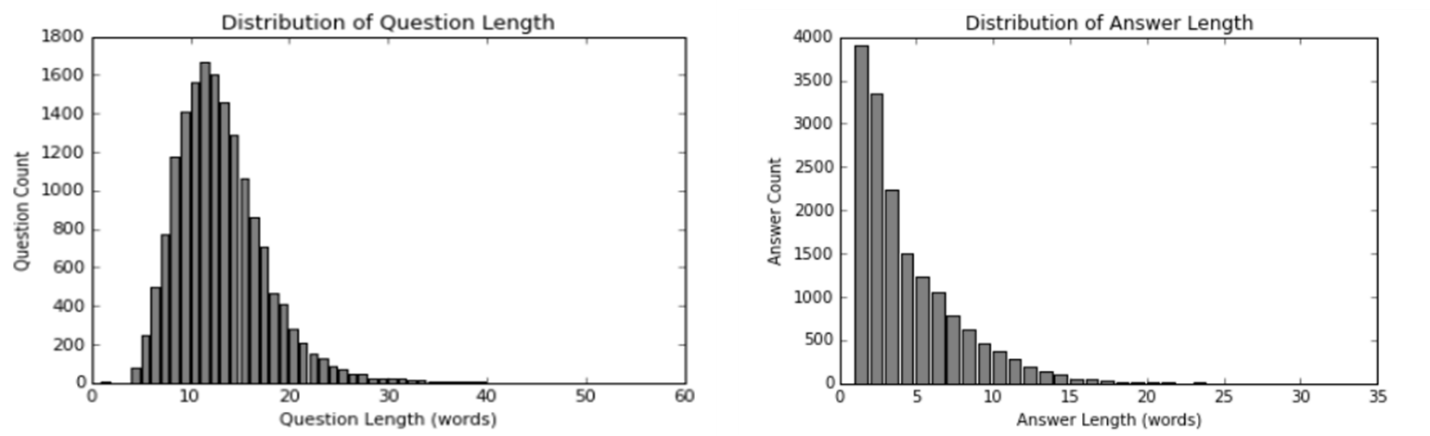
While the dataset may be somewhat smaller than other offerings it brings several key benefits. As aforementioned, the first benefit is natural language questions that reflect the way people communicate with one another. By comparing the distribution of the length of questions for each dataset, one can see that the Classics Comprehension dataset, along with the SQUAD and TriviaQA dataset that both also have natural language questions, has an average around thirteen words per question (see figure 2). In MS MARCO and WikiSuggest the average words per question is much lower; only six words and five words respectively reflecting the influence of the traditional keyword query nature seen in search engine generated questions (see figure 3). (Eunsol,’17) (Nguyen, ‘16)

Figure 2: Classics Comprehension Question Distribution

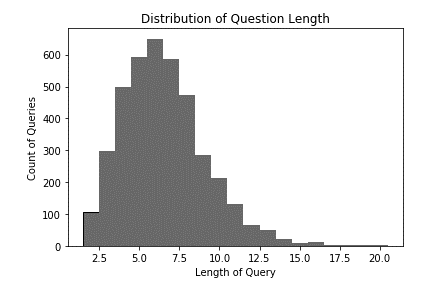


Figure 3: MS MARCO question distribution

Another key differentiator is the longer context passages compared to the SQUAD and Children’s Book datasets where the questions were based on paragraph level passages (an average of about one hundred fifty words) and sentence level respectively (Hill, ‘16) (Gunel). The MS MARCO dataset combines between six to ten paragraph passages from different sources and has a slightly longer total word count with a bimodal distribution that peaks around five hundred fifty words and a thousand words. Some of the best prior work with longer range passages can be seen in the TriviaQA dataset where each question is answered based on six supporting documents with an average length of 2,895 context words (Nguyen,‘16)(Mandar, ‘17). Quickly spot checking several examples in the Trivia QA dataset illustrates that several of them had answers in the first paragraph or even the first sentence. This introduces a problematic bias for understanding longer range text documents and solving the vanishing gradient problem. Swayamdipta et al. mention the most successful reading comprehension models trim the context to the first eight hundred words. Anyone that has fallen asleep reading one of the classic books for an English homework assignment can attest to the importance of finishing the entire chapter this highlights the third key benefit of the Classics Comprehension dataset that the questions related to various places throughout the text. The average length of the Classics Comprehension context is over five thousand words which also corresponds to the average length of a long research paper (see figure 4). Fully understanding bodies of academic papers, such as medical or legal applications, are especially important given that the results and error analysis are at the end of the report further emphasizing the value of successfully comprehending longer range datasets.

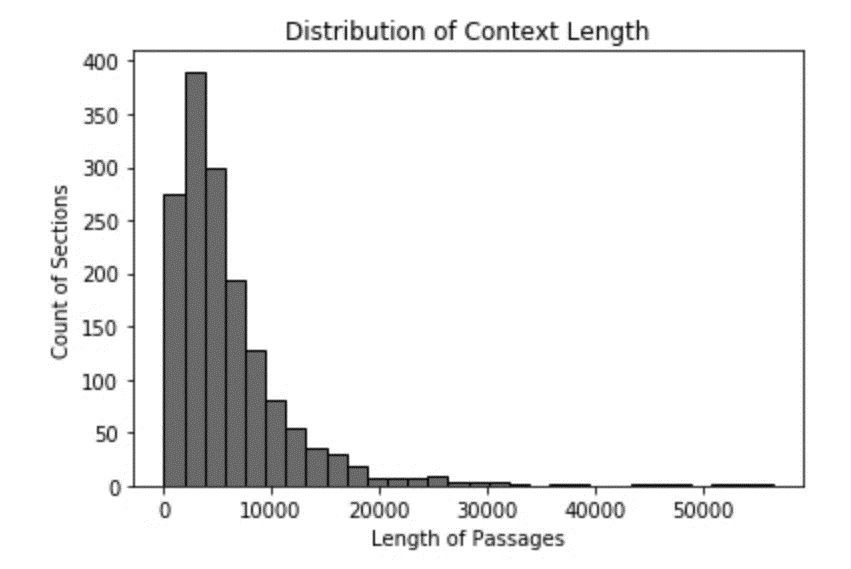


Figure 4: Classic Comprehension

The new medium provided by the Classics Comprehension dataset also facilitates an easy complexity analysis of the body of text the computer is working to understand. Most popular and classic literature received a Lexile Score that measures the semantic, syntactic, and vocabulary features used. The Gutenberg books are generally at the higher end of the Lexile spectrum which ranges from 300 to 1700 (see figure 4) (Lexile). Another important benefit is that this new dimension enables a deeper understanding of how the model performs against more complex texts and offers a new benchmark against human intelligence as age and academic grade bands are associated with the Lexile scores.

Figure 5: Classic Comprehension Lexile Score

For a final measure of analysis, we turned our attention to the questions asked about the datasets. Without knowing the location of the answers in the book chapters, it was impossible to replicate the syntactic divergence that the SQUAD and Trivia QA papers used in their data analysis section so instead we aligned with MS MARCO approach. The evaluation required tagging the question based on the seven commonly used keywords in questions (who, what, when, where, how, why and which). The results shown in table 2 highlight the bias associated with several of the data sources. The Classics Comprehension questions have a disproportionately high number of ‘who’ questions, but this makes sense given the importance of recalling characters in understanding books. The MS MARCO dataset also reveals it has bias towards ‘how’ and ‘where’ based questions with 27% and 12% allocated those categories; reflecting the frequent use of internet searches for ‘how to’ accomplish various tasks and ‘where to’ find entities. The SQUAD dataset has synthetically generated questions and falls somewhere between the two extremes. In summary, it is important to include a variety of data sources to obtain a more comprehensive question answering system. (Nguyen, ’16)( Gunel)

Table 2: Question Analysis Across datasets

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Question Type | CC | CC | MS MARCO[[1]](#footnote-1) | MS MARCO Rebalanced | SQUAD[[2]](#footnote-2) |
| What | 7,181 | 44% | 42% | 61% | 55% |
| Who | 2,915 | 18% | 2% | 3% | 12% |
| How | 1,742 | 11% | 15% | 27% | 13% |
| When | 1,573 | 10% | 2% | 5% | 8% |
| Why | 1,094 | 7% | 2% | 5% | 2% |
| Where | 921 | 6% | 4% | 12% | 5% |
| Which | 593 | 4% | 1% | 4% | 5% |
| Other | 445 | 3% | 31% | NA | 0% |

## **5 Results & Discussion**

We developed several logistic regression Bag of Words (BoW) models and compared their accuracy with that of two LSTM models. Logistic regression was selected to align with the approach in the SQUAD paper. In developing the initial baseline model we ran into challenges with long computational run time and low accuracy (weighted F1 score) that can be attributed to the context length that more than doubles prior datasets and significantly exceeds the SQUAD length (Model 1 results shown in table 3). This posed a significant challenge as the advanced algorithms (CNN, RNN, LSTM, Attention or BiDAF) would add computational time and corresponding cost increases. Evaluating approaches on other longer datasets, it became apparent that longer training was unrealistic given the project timeline and resources. For example, Swayamdipta wrote that it took ‘2-3 days to train on a single NVIDIA P100 GPU’ and they were working with the Trivia QA dataset which has passages that are half as long as the Classics Comprehension dataset (Swayamdipta et al., ’17). Setting aside the issue that the P100 was not an easily accessible option in Google Cloud, it also would potentially result in a cost of $70 for a model run—assuming everything went smoothly on that run and there were no amendments needed—that would still be too expensive and too slow.

The most recent research on reading comprehension techniques emphasize the importance of identifying key places in the text that contain the answer to the question and encouraging the computer to more heavily analyze just that section (Eunsol,’17) (Swayamdipta et al., ’17). After ruling out more computationally intensive techniques, we needed a simple heuristic to identify key passages. Each question was compared to each sentence in the text using a cosine similarity calculation to pick the top sentence. To focus on keywords TF-IDF was leveraged, but it was unclear how much context to use when creating the vector. Selecting the correct passage is a difficult task because the location of answer spans is unavailable making it an unsupervised task. Ultimately, we decided to try testing both a vector that incorporates the entire chapter and one that compared one sentence to one question at a time. Keeping all other parameters consistent, as table 3 shows, the performance was extremely similar across both context weighting approaches. Calculating the cosine distance for one sentence and one question provided a slight increase in accuracy: 0.18% vs 0.21% in the Bag of Words approach and 16.0% vs 16.3% for the LSTM approach, but the accuracy came with some additional computation time. Using such a simple information retrieval pipeline reduced the runtime from several hours to minutes while increasing the model accuracy by ~20 fold.

Table 3: Model Results

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Model** | **Accuracy** | **Time** |
| 1 | Bag of Words (BoW) LR, all data | 0.01% | 10+ hrs. 20 CPU 40GB |
| 2 | BoW, top sentence, LR, TF-IDF Chapters | 0.18% | 12 min 8CPU 16 GB |
| 3 | BoW, top sentence, LR, TF-IDF 1:1 | 0.21% | <30 min |
| 4 | BoW, top sentence, LR, TF-IDF Chapters  Binary | 7.35% | 6 min  8CPU 16 GB |
| 5 | LSTM, top sentence, TF-IDF Chapters, Binary | 16.0% | 2h 50 min  8CPU 16 GB |
| 6 | LSTM, top sentence, 1:1, Binary | 16.3% | 2h 50 min  8CPU 16 GB |

After identifying key sentences, we implemented the LSTM code from the BabI Children’s book test because that algorithm does very well with sentences. The initial epochs perform very well with a 16% accuracy rate, but then the LSTM model begins to overfit as evidenced by the accuracy decreasing to 13% with an increased number of epochs. This accuracy jump initially looks impressive, but it is overstated. As part of implementing the BaBI code, the answers were reduced to a single word. To accomplish this the median value was selected from the encoded answer words. When the same simplification is applied to the BoW model the accuracy jumps to 7% and the compute time is cut in half. These models demonstrate that the run time and accuracy of any question and answer dataset is highly impacted by the number of words being predicted, not just the amount of words in the vocabulary. The Classics Comprehension dataset has an average of four words per answer so a higher model accuracy is definitely possible with some more complex models better suited to such inputs. Additionally, the Children’s Book test, SQUAD and the Trivia QA dataset all initially had accuracy scores around fifty percent; further supporting the notion that improvements can be made.

## **6 Conclusion**

After optimizing the sentence selection technique, there is a great opportunity to concatenate the top 3-10 sentences and build more complex models based upon the simplified information retrieval pipeline. Prior research suggests that a sliding window model, a recurrent neural network, and an attention model would be ideal next steps and offer strong benchmarks to well conducted research.

Finally, the new dataset has the potential to expand. Back of the envelope calculations suggest that the BookRags website has thirty-five million questions on one hundred ninety thousand books. In the future it would be interesting to leverage the Lexile scores to build Bayesian Hierarchical models that learn in a compounding fashion which mirrors how people are taught reading comprehension.

## **7 References**

CLASSICS COMPREHENSION:

“Project Gutenberg.” *Project Gutenberg*, www.gutenberg.org/.

*BookRags*, BookRags, www.bookrags.com/.

“The Lexile Framework for Reading.” *Lexile*, lexile.com/.

Reeve, Jonathan. “JonathanReeve/Chapterize.” *GitHub*, github.com/JonathanReeve/chapterize/tree/master/chapterize.

^Didn’t save much time because chapters are frequently bunched at random intervals so I ended up breaking up the text manually ☹

https://github.com/jphilippou27/nlp\_final\_project

PAPERS:

Eunsol, Choi, et al. “Hierarchical Question Answering for Long Documents.” *[1611.01839] Hierarchical Question Answering for Long Documents*, 8 Feb. 2017, arxiv.org/abs/1611.01839.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The Goldilocks principle: Reading children’s books with explicit memory representations, 2015. URL http://arxiv.org/ abs/1511.02301. arXiv:1511.02301.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In Proceedings of the Conference of the Empirical Methods in Natural Language Processing.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text, 2016. arXiv:1606.05250.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proc. of ACL, 2017.

Swayamdipta, Swabha, et al. “Multi-Mention Learning for Reading Comprehension with Neural Cascades.” *[1711.00894] Multi-Mention Learning for Reading Comprehension with Neural Cascades*, 2 Nov. 2017, arxiv.org/abs/1711.00894.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In Workshop in Advances in Neural Information Processing Systems. https://arxiv.org/pdf/1611.09268.pdf.

OTHER:

Eunsol, Choi. Scaling up Reading Comprehension. Nov. 2017. https://nlp.stanford.edu/seminar/details/echoi.pdf

“Graphics Processing Unit (GPU) | Google Cloud.” Google, Google, cloud.google.com/gpu/.

Gunel, Beliz. Question Answering on SQUAD. web.stanford.edu/class/cs224n/reports/6880391.pdf.

keras-team. “Keras-Team/Keras.” GitHub, github.com/keras team/keras/blob/master/examples/babi\_memnn.py.

Q&A DATA SOURCES:

<https://rajpurkar.github.io/SQuAD-explorer/>

<http://nlp.cs.washington.edu/triviaqa/index.html>

<http://www.msmarco.org/>

<https://research.fb.com/downloads/babi/>

## **Appendix**

#### Texts Consulted

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | Little Women | 51 | The Autobiography of Benjamin Franklin |
| 2 | The Wonderful Wizard of Oz | 52 | The Awakening |
| 3 | The Adventures of Huckleberry Finn | 53 | Candide |
| 4 | Alice’s Adventures in Wonderland | 54 | Walden |
| 5 | Peter Pan | 55 | Romeo and Juliet |
| 6 | The Wind in the Willows | 56 | Pygmalion |
| 7 | Anne of Green Gables | 57 | The Story of my Life |
| 8 | The Secret Garden | 58 | Ethan Frome |
| 9 | A Christmas Carol | 59 | On the Origin of Species |
| 10 | Black Beauty | 60 | Uncle Tom's Cabin |
| 11 | Oliver   Twist | 61 | The Legend of Sleepy Hollow |
| 12 | Dracula | 62 | The Turn of the Screw |
| 13 | David   Copperfield | 63 | The Scarlet Letter |
| 14 | Treasure   Island | 64 | Persuasion |
| 15 | Around   the World in Eighty Days | 65 | An Occurrence at Owl Creek Bridge |
| 16 | Time   Machine | 66 | War of the Worlds |
| 17 | Adventures   of Sherlock Holmes | 67 | Call of the Wild |
| 18 | Jungle   Book | 68 | Three Men in a Boat |
| 19 | Frankenstein | 69 | Northanger Abbey |
| 20 | Gulliver’s   Travels | 70 | Mansfield Park |
| 21 |  | 71 | Notes from the Underground |
| 22 | Pride and Prejudice | 72 | Don Quixote |
| 23 | The Importance of Being Earnest | 73 | An Ideal Husband |
| 24 | A Tale of Two Cities | 74 | Madame Bovary |
| 25 | The Metamorphosis | 75 | Phantom of the Opera |
| 26 | Moby Dick | 76 | The Prince and the Pauper |
| 27 | Dr. Jekyll and Mr. Hyde | 77 | Middlemarch |
| 28 | The Adventures of Tom Sawyer | 78 | Paradise Lost |
| 29 | Anne of the Island | 79 | North and South |
| 30 |  | 80 | The Age of Innocence |
| 31 | The Prince | 81 | Bleak House |
| 32 | Great Expectations | 82 | A portrait of the Artist as a Young Man |
| 33 | War and Peace | 83 | The Fall of the House of Usher |
| 34 | The Picture of Dorian Gray | 84 | The Odyssey |
| 35 | Jane Eyre | 85 | Ivanhoe |
| 36 | Emma | 86 | Hamlet |
| 37 | The Iliad | 87 | The Trial |
| 38 | Vanity Fair | 88 | A Room with A View |
| 39 | Leviathan | 89 | The Secret Agent |
| 40 | Wuthering Heights | 90 | The Three Musketeers |
| 41 | Dubliners | 91 | Vanity Fair |
| 42 | The Count of Monte Cristo | 92 | The Woman in White |
| 43 | The Souls of Black Folk | 93 | Tarzan of the Apes |
| 44 | Narrative of the Life of Frederick Douglass | 94 | Twelve Years a Slave |
| 45 | My Antonia | 95 | The Idiot |
| 46 | The Brothers Karamazov | 96 | The House of Mirth |
| 47 | Sense and Sensibility | 97 | This Side of Paradise |
| 48 | Beyond Good and Evil | 98 | The Canterbury Tales |
| 49 | Les Miserables | 99 | Macbeth |
| 50 | The Hounds of Baskervilles |  |  |
|  |  |  |  |
| Backup | Herland | Backup | Our Mutual Friend |
| Backup | The Canterbury Tales | Backup | Anne of the Island |
| Backup | Lady WinderMere's Fan | Backup | The Grand Inquisitor |
| Backup | The Waste Land | Backup | The Canterville Ghost |
| Backup | My Man Jeeves | Backup | The Scarlet Pimpernel |
| Backup | Maggie A girl of the streets | Backup | Villette |
| Backup | Roughing it | Backup |  |
| Backup | The Man that Corrupted Hadleyburg | Backup | Lord Jim |
| Backup | The Monkey's Paw | Backup | Othello |
| Backup | A Midsummer Night's Dream | Backup | Nicholos Nickleby |

-To do: clean up book list

-show impact of answer length on accuracy for BoW

- Mulitword answers for LSTM

-RNN & Attention!

1. Taken from original paper [↑](#footnote-ref-1)
2. Taken from subsequent paper, not original authors, also only a sample of the data [↑](#footnote-ref-2)