

BookQA: Stories of Challenges and Opportunities

Stefanos Angelidis^{1*} Lea Frermann² Diego Marcheggiani³ Roi Blanco³ Lluís Màrquez³

¹Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh

²School of Computing and Information Systems, The University of Melbourne

³Amazon Research

s.angelidis@ed.ac.uk lea.frermann@unimelb.edu.au

{marchegg, roiblan, lluismv}@amazon.com

Abstract

We present a system for answering questions based on the full text of books (BookQA), which first selects book passages given a question at hand, and then uses a memory network to reason and predict an answer. To improve generalization, we pretrain our memory network using artificial questions generated from book sentences. We experiment with the recently published NarrativeQA corpus, on the subset of *Who* questions, which expect book characters as answers. We experimentally show that BERT-based retrieval and pretraining improve over baseline results significantly. At the same time, we confirm that NarrativeQA is a highly challenging data set, and that there is need for novel research in order to achieve high-precision BookQA results. We analyze some of the bottlenecks of the current approach, and we argue that more research is needed on text representation, retrieval of relevant passages, and reasoning, including commonsense knowledge.

1 Introduction

Considerable volume of research work has looked into various Question Answering (QA) settings, ranging from retrieval-based QA (Voorhees, 2001) to recent neural approaches that reason over Knowledge Bases (KB) (Bordes et al., 2014), or raw text (Shen et al., 2017; Deng and Tam, 2018; Min et al., 2018). In this paper we use the NarrativeQA corpus (Kocisky et al., 2018) as a starting point and focus on the task of answering questions from the full text of books, which we call BookQA. BookQA has unique characteristics which prohibit the direct application of current QA methods. For instance, (a) books are usually orders of magnitude longer than the short texts (e.g.,

Wikipedia articles) used in neural QA architectures; (b) many facts about a book story are never made explicit, and require external or commonsense knowledge to infer them; (c) the QA system cannot rely on pre-existing KBs; (d) traditional retrieval techniques are less effective in selecting relevant passages from self-contained book stories (Kocisky et al., 2018); (e) collecting human-annotated BookQA data is a significant challenge; (f) stylistic disparities in the language used among different books may hinder generalization.

Additionally, the style of book questions may vary significantly, with different approaches being potentially useful for different question types: from queries about story facts that have entities as answers (e.g., *Who* and *Where* questions); to open-ended questions that require the extraction or generation of longer answers (e.g., *Why* or *How* questions). The difference in reasoning required for different question types can make it very hard to draw meaningful conclusions.

For this reason, we concentrate on the task of answering *Who* questions, which expect book characters as answers (e.g., “*Who is Harry Potter’s best friend?*”). This task allows to simplify the output and evaluation (we look for entities, and we can apply precision-based and ranking evaluation metrics), but still retains the important elements of the original NarrativeQA task, i.e., the need to explore over the full content of the book and to reason over a deep understanding of the narrative. Table 1 exemplifies the diversity and complexity of *Who* questions in the data, by listing a set of questions from a single book, which require increasingly complex types of reasoning.

NarrativeQA (Kocisky et al., 2018) is the first publicly available dataset for QA over long narratives, namely the full text of books and movie scripts. The full-text task has only been addressed

*Work done while first author was interning at Amazon.

<i>Who is Emily in love with?</i>
<i>Who is Emily imprisoned by?</i>
<i>Who helps Emily escape from the castle?</i>
<i>Who owns the castle in which Emily is imprisoned?</i>
<i>Who became Emily's guardian after her father's death?</i>

Table 1: *Who* questions from NarrativeQA for the book *The Mysteries of Udolpho*, by Ann Radcliffe. The diversity and complexity of questions in the corpus remains high, even when considering only the subset of *Who* questions that expect characters as answers.

by Tay et al. (2019), who proposed a curriculum learning-based two-phase approach (*context selection* and *neural inference*). More papers have looked into answering NarrativeQA’s questions from only book/movie *summaries* (Indurthi et al., 2018; Bauer et al., 2018; Tay et al., 2018a,b; Nishida et al., 2019). This is a fundamentally simpler task, because: i) the systems need to reason over a much shorter context, i.e., the summary; and ii) there is the certainty that the answer can be found in the summary. This paper is another step in the exploration of the full NarrativeQA task, and embraces the goal of finding an answer in the complete book text. We propose a system that first selects a small subset of relevant book passages, and then uses a memory network to reason and extract the answer from them. The network is specifically adapted for generalization across books. We analyze different options for selecting relevant contexts, and for pretraining the memory network with artificially created question–answer pairs. Our key contributions are: i) this is the first systematic exploration of the challenges in full-text BookQA, ii) we present a full pipeline framework for the task, iii) we publish a dataset of *Who* questions which expect book characters as an answer, and iv) we include a critical discussion on the shortcomings of the current QA approach, and we discuss potential avenues for future research.

2 Book Character Questions

NarrativeQA was created using a large annotation effort, where participants were shown a human-curated *summary* of a book/script and were asked to produce question-answer pairs *without referring to the full story*. The main task of interest is to answer the questions by looking at the *full story* and not at the summary, thus ensuring that answers cannot be simply copied from the story. The full corpus contains 1,567 stories (split equally between books and movies) and 46,765 questions.

We restrict our study to *Who* questions about books, which have *book characters* as answers (e.g., “*Who is charged with attempted murder?*”). Using the book preprocessing system, book-nlp (see Section 3.1), and a combination of automatic and crowdsourced efforts, we obtained a total of 3,427 QA pairs, spanning 614 books.¹

3 BookQA Framework

The length of books and limited annotated data prohibit the application of end-to-end neural QA models that reason over the full text of a book. Instead, we opted for a pipeline approach, whose components are described below.

3.1 Book & Question Preprocessing

Books and questions are preprocessed in advance using the book-nlp parser (Bamman et al., 2014), a system for character detection and shallow parsing in books (Iyyer et al., 2016; Frermann and Szarvas, 2017) which provides, among others: sentence segmentation, POS tagging, dependency parsing, named entity recognition, and coreference resolution. The parser identifies and clusters character mentions, so that all coreferent (direct or pronominal) character mentions are associated with the same unique character identifier.

3.2 Context Selection

In order to make inference over book text tractable and give our model a better chance at predicting the correct answer, we must restrict the context to only a small number of book sentences. We developed two context selection methods to retrieve relevant book passages, which we define as windows of 5 consecutive sentences:

IR-style selection (BM25F): We constructed a searchable *book index* to store individual book sentences. We replace every book character mention, including pronoun references, with the character’s unique identifier. At retrieval time, we similarly replace character mentions in each question, and rank passages from the corresponding book using BM25F (Zaragoza et al., 2004).

BERT-based selection: We developed a neural context selection method, based on the BERT language representation model (Devlin et al., 2019). A pretrained BERT model is fine-tuned to predict

¹To obtain the BookQA data, follow the instructions at: <https://github.com/stangelid/bookqa-who>.

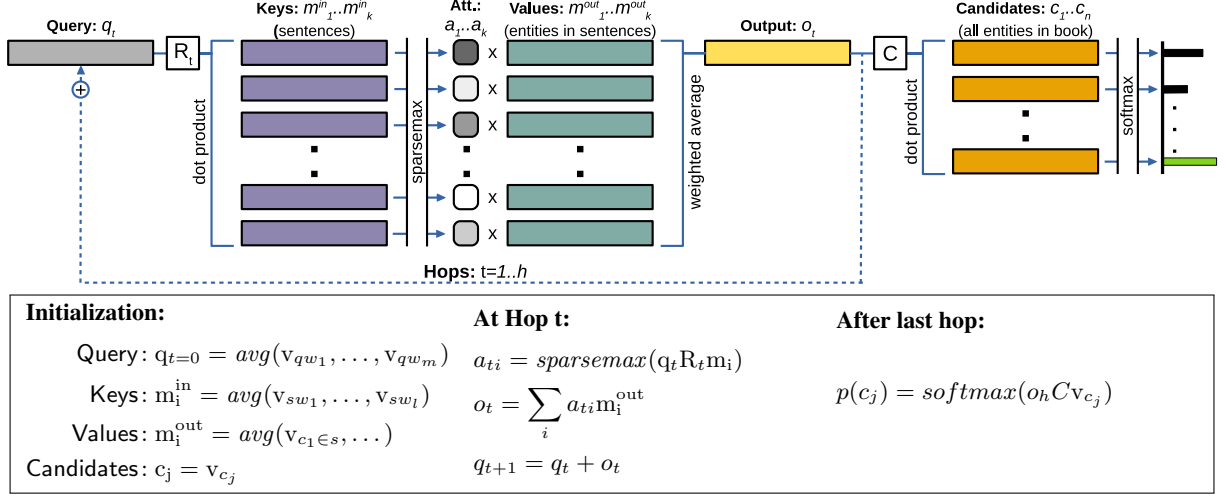


Figure 1: Overview of our Key-Value Memory Network for BookQA. Encodings of questions, keys (selected sentences), and values (characters mentioned in those sentences) are loaded. After multiple hops of inference, the model’s output is compared against the candidate answers’ encodings to make a prediction.

if a sentence is relevant to a question, using positive (*questions, summary sentence*) training pairs which have been heuristically matched. Randomly sampled negative pairs were also used. At retrieval time, a question is used to retrieve relevant passages from the full text of a book.

3.3 Neural Inference

Having replaced character mentions in questions and books with character identifiers, we first pre-train word2vec embeddings (Mikolov et al., 2013) for all words and book characters in our corpus.² Our neural inference model is a variant of the Key-Value Memory Network (KV-MemNet) (Miller et al., 2016), which has been previously applied to QA tasks over KBs and short texts. The original model was designed to handle a fixed set of potential answers across all QA examples, as do most neural QA architectures. This comes in contrast with our task, where the pool of candidate characters is different for each book. Our KV-MemNet variant, illustrated in Figure 1, uses a dynamic output layer where different candidate answers are made available for different books, while the remaining model parameters are shared.

A question is initially represented as q_0 , i.e., the average of its word embeddings³ (gray vector). The Key memories $m_1^{\text{in}} \dots m_k^{\text{in}}$ (purple vectors) are filled with the k most relevant sentences, as retrieved from the context selection step, us-

ing the average of their word embeddings. Value memories $m_1^{\text{out}} \dots m_k^{\text{out}}$ (green vectors) contain the average embedding of all characters mentioned in the respective sentence, or a padding vector if no character is mentioned. Candidate embeddings $c_1 \dots c_n$ (orange vectors) hold the embeddings of every character in the current book. The model makes multiple reasoning hops $t = 1 \dots h$ over the memories. At each hop, q_t is passed through linear layer R_t and is then compared against all key memories. The *sparsemax*-normalized (Martins and Astudillo, 2016) attention weights $a_1 \dots a_k$ are then used for obtaining output vector o_t , as the weighted average of value memories. The process is repeated h times, and the final output is passed through linear layer C , before being compared against all candidate vectors via dot-product, to obtain the final prediction. The model is trained using negative log-likelihood.

3.4 Pretraining

A significant obstacle towards effective BookQA is the limited amount of data available for supervised training. A potential avenue for overcoming this is pretraining the neural inference model on an auxiliary task, for which we can generate orders of magnitude more training examples. To this end, we generated 688,228 artificial questions from the book text using a set of simple pruning rules over the dependency trees of book sentences. We used all book sentences where a character mention is the agent or the patient of an active voice verb, or the patient of a passive voice verb. Two examples

²Character identifiers are treated like all other tokens.

³Experiments with more sophisticated question/sentence representation variants showed no significant improvements.

Metric →	P@1		P@5		MRR	
Context selection →	BM25F	BERT	BM25F	BERT	BM25F	BERT
Baselines:						
Book frequency		15.73		56.29		0.337
Context frequency	10.53	13.80	51.42	53.02	0.276	0.305
KV-MemNet:						
No pretraining	15.57±0.97	15.89±0.95	58.18±1.57	58.77±1.29	0.339±0.006	0.343±0.008
Pretrain w/ Artif. Qs	15.92±0.73	18.73±1.07	61.25±0.74	62.81±1.07	0.351±0.005	0.376±0.006

Table 2: Precision scores (P@1, P@5), and Mean Reciprocal Rank (MRR) for frequency-based baselines and our system, with and without pretraining. We report average and standard deviation over 50 runs.

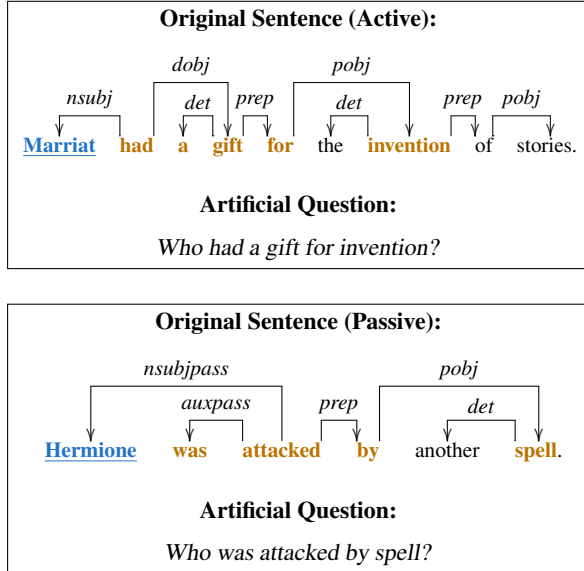


Figure 2: Examples of artificial questions generated from the dependency trees of an active voice (top) and a passive voice (bottom) sentence. The correct answer (verb’s subject) is marked with **blue**, whereas the **yellow** words are used in the question. The remaining words are discarded by pruning the dependency tree.

are illustrated in Figure 2: at the top, the active voice sentence “Marriat had a gift for the invention of stories.” is transformed into the question “Who had a gift for invention?” and, at the bottom, the passive voice sentence “Hermione was attacked by another spell.” is transformed into the question “Who was attacked by a spell?”. The previous 20 book sentences, including the source sentence, are used as context during pretraining.

4 Experimental Setup

For every question, 100 sentences (top 20 passages of five sentences) were selected as contexts using our retrieval methods. We used word and book character embeddings of 100 dimensions. The number of reasoning hops was set to 3. When no pretraining was performed, we trained on the real QA examples for 60 epochs, using Adam with ini-

tial learning rate of 10^{-3} , which we reduced by 10% every two epochs. Word and character embeddings were fixed during training. When using pretraining, we trained the memory network for one epoch on the auxiliary task, including the embeddings. Then, the model was fine-tuned as described above on the real QA examples where, again, embeddings were fixed. We use Precision at the 1st and 5th rank (P@1 and P@5) and Mean Reciprocal Rank (MRR) as evaluation metrics. We adopted a 10-fold cross validation approach and performed 5 trials for each cross validation split, for a total of 50 experiments.

Baselines: We implemented a random baseline and two frequency-based baselines, where the most frequent character in the entire book (Book frequency) or the selected context (Context frequency) was selected as the answer.

5 Results

Our main results are presented in Table 2. Firstly, we observe one of the dataset’s biases, as the book’s most frequent character is the correct answer in more than 15% of examples, whereas selecting a character at random would only yield the correct answer 2.5% of the time.

With regards to our BookQA pipeline, the results confirm that BookQA is a very challenging task. Without pretraining, our KV-MemNet which uses IR contexts achieves 15.57% P@1, and it is slightly outperformed by its BERT-based counterpart.⁴ When pretraining the memory network with artificial questions, the BERT-based model achieves 18.73% P@1. The same trend is observed with the other metrics.

Number of hops: We also calculated the impact of the number of hops with respect to the P@1 for a pretrained model fine-tuned with BERT-selected

⁴Despite the similar performance to the Book frequency baseline, we *did not* observe that our model was systematically selecting the most frequent character as the answer.

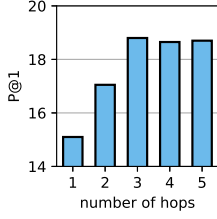


Figure 3: P@1 for different number of hops.

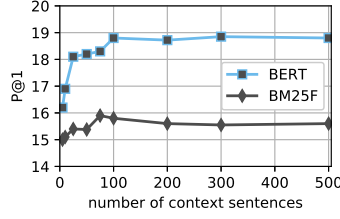


Figure 4: P@1 for varying context sizes from BM25F and BERT.

contexts. Figure 3 shows that performance increases up to 3 hops and then it stabilizes.

Context size: We expected the context size (i.e., the number of retrieved sentences that we store in the memory slots of our KV-MemNet) to significantly affect performance. Smaller contexts, obtained by only retrieving the topmost relevant passages, might miss important evidence for answering a question at hand. Conversely, larger contexts might introduce noise in the form of irrelevant sentences that hinder inference. Figure 4 shows the performance of our method when varying the number of context sentences (or, equivalently, memory slots). The neural inference model struggles for very small context sizes and achieves its best performance for 75 and 100 context sentences obtained by BM25F and BERT, respectively. For both alternatives, we observe no further improvements for larger contexts.

Pretraining size & epochs: A key component of our BookQA framework is the pretraining of our neural inference model with artificially generated questions. Although it helped achieve the highest percentage of correctly answered questions, the performance gains were relatively small given the number of artificial questions used to pretrain the model. We further investigated the effect of pretraining by varying the number of artificial questions used during training and the number of pretraining epochs. Figure 5 shows the QA performance achieved on the real BookQA questions (using BM25F or BERT contexts) after pretraining on a randomly sampled subset of the artificial questions. For our BERT-based variant, the percentage of correctly answered questions increases steadily, but flattens out when reaching 75% of pretraining set usage. On the contrary, when using BM25F contexts we achieved insignificant gains, with performance appearing constrained by the quality of retrieved passages. In Figure 6 we show

correct character mentioned in context	BM25F	69.7%
	BERT	74.7%
full evidence found in context		27%
partial evidence found in context	BM25F	47%
no evidence found in context		26%

Table 3: Percentage of contexts where the correct character is mentioned (top). Percentage of contexts where full/partial/no evidence for the answer was found according to crowd-workers who examined a sample of 100 cases (bottom).

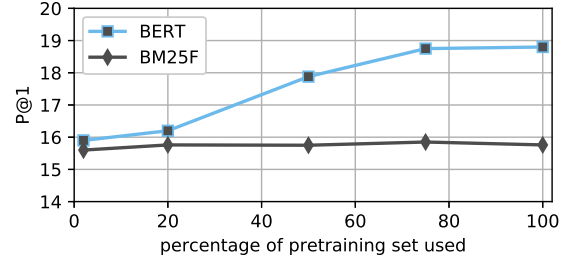


Figure 5: P@1 for varying percentage of pretraining questions used (BM25F and BERT contexts).

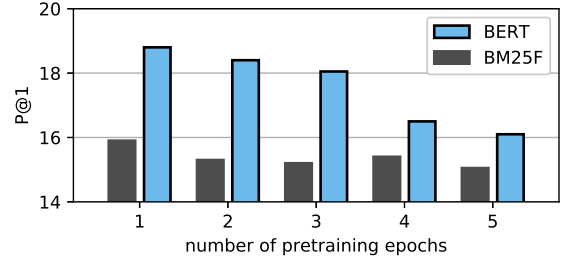


Figure 6: P@1 as a function of pretraining epochs for BM25F and BERT contexts.

P@1 scores as a function of the number of pretraining epochs. Best performance is achieved after only one epoch for both variants, indicating that further pretraining might cause the model to overfit to the simpler type of reasoning required for answering artificial questions.

5.1 Further Discussion

Despite the limitation to *Who* questions, the employment of strong models for context selection and neural inference, and our pretraining efforts, the overall BookQA accuracy remains modest, as our best-performing system achieves a P@1 score below 20%. Even when we only allowed our system to answer if it was very confident (according to the probability difference between top-ranked candidate answers), it answered correctly 35% of times.

We have identified a number of reasons which inhibit better performance. Firstly, the passage selection process constrains the answers that can be logically inferred. We provide our findings in regards to this claim in Table 3. We calculated that the correct answer appears in the IR-selected contexts in 69.7% of cases. For BERT-selected contexts it appears in 74.7% of cases. In practice, however, these upper-bounds are not achievable; even when the correct answer appears in the context, there is no guarantee that enough evidence exists to infer it. To further investigate this, we ran a survey on Amazon Mechanical Turk, where participants were asked to indicate if the selected context (IR-retrieved) contained partial or full evidence for answering a question. For a set of 100 randomly sampled questions, participants found full evidence for answering a question in just 27% of cases. Only partial evidence was found in 47% of cases, and no evidence in the remaining 26%.

Manual inspection of context sentences indicated that a common reason for the absence of full evidence is the inherent vagueness of literary language. Repeated expressions or direct references to character names are often avoided by authors, thus requiring very accurate paraphrase detection and coreference resolution. We believe that commonsense knowledge is particularly crucial for improving BookQA. When exploring the output of our system, we repeatedly found cases where the model failed to arrive at the correct answer due to key information being left implicit. Common examples we identified were: i) character relationships which were clear to the reader, but never explicitly described (e.g., “*Who did Mark’s best friend marry?*”); ii) the attitude of a character towards an event or situation (e.g., “*Who was angry at the school’s policy?*”); iii) the relative succession of events (e.g., “*Who did Marriat talk to after the big fight?*”). The injection of commonsense knowledge into a QA system is an open problem for general and, consequently, BookQA.

In regards to pretraining, the lack of further improvements is likely related to the difference in the type of reasoning required for answering the artificial questions and the real book questions. By construction, the artificial questions will only require that the model accurately matches the source sentence, without the need for complex or multi-hop reasoning steps. In contrast, real book questions require inference over information spread across

many parts of a book. We believe that our proposed auxiliary task mainly helps the model by improving the quality of word and book character representations. It is, however, clear from our results that pretraining is an important avenue for improving BookQA accuracy, as it can increase the number of training instances by many orders of magnitude with limited human involvement. Future work should look into automatically constructing auxiliary questions that better approximate the types of reasoning required for realistic questions on the content of books.

We argue that the shortcomings discussed in previous paragraphs, i.e., the lack of evidence in retrieved passages, the difficulty of long-term reasoning, the need for paraphrase detection and commonsense knowledge, and the challenge of useful pretraining, are not specific to *Who* questions. On the contrary, we expect that the requirement for novel research in these areas will generalize or, potentially, increase in the case of more general questions (e.g., open-ended questions).

6 Conclusions

We presented a pipeline BookQA system to answer character-based questions on NarrativeQA, from the full book text. By constraining our study to *Who* questions, we simplified the task’s output space, while largely retaining the reasoning challenges of BookQA, and our ability to draw conclusions that will generalize to other question types. Given a *Who* question, our system retrieves a set of relevant passages from the book, which are then used by a memory network to infer the answer in multiple hops. A BERT-based trained retrieval system, together with the usage of artificial question-answer pairs to pretrain the memory network, allowed our system to significantly outperform the lexical frequency-based baselines. The use of BERT-retrieved contexts improved upon a simpler IR-based method although, in both cases, only partial evidence was found in the selected contexts for the majority of questions. Increasing the number of retrieved passages did not result in better performance, highlighting the significant challenge of accurate context selection. Pretraining on artificially generated questions provided promising improvements, but the automatic construction of realistic questions that require multi-hop reasoning remains an open problem. These results confirm the difficulty of the BookQA chal-

lenge, and indicate that there is need for novel research in order to achieve high-quality BookQA. Future work on the task must focus on several aspects of the problem, including: (a) improving context selection, by combining IR and neural methods to remove noise in the selected passages, or by jointly optimizing for context selection and answer extraction (Das et al., 2019); (b) using better methods for encoding questions, sentences, and candidate answers, as embedding averaging results in information loss; (c) pretraining tactics that better mimic the real BookQA task; (d) incorporation of commonsense knowledge and structure, which was not addressed in this paper.

Acknowledgments We would like to thank Hugo Zaragoza and Alex Klementiev for their valuable insights, feedback and support on the work presented in this paper.

References

- David Bamman, Ted Underwood, and Noah A. Smith. 2014. [A bayesian mixed effects model of literary character](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379. Association for Computational Linguistics.
- Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. [Commonsense for generative multi-hop question answering tasks](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.
- Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. [Question answering with subgraph embeddings](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 615–620. Association for Computational Linguistics.
- Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. [Multi-step retriever-reader interaction for scalable open-domain question answering](#). In *ICLR 2019*.
- Haohui Deng and Yik-Cheung Tam. 2018. [Read and comprehend by gated-attention reader with more belief](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 83–91, New Orleans, Louisiana, USA. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Lea Frermann and György Szarvas. 2017. [Inducing semantic micro-clusters from deep multi-view representations of novels](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1873–1883. Association for Computational Linguistics.
- Sathish Reddy Indurthi, Seunghak Yu, Seohyun Back, and Heriberto Cuayáhuitl. 2018. [Cut to the chase: A context zoom-in network for reading comprehension](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 570–575, Brussels, Belgium. Association for Computational Linguistics.
- Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. [Feuding families and former friends: Unsupervised learning for dynamic fictional relationships](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544. Association for Computational Linguistics.
- Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018. [The narrativeqa reading comprehension challenge](#). *Transactions of the Association for Computational Linguistics*, 6:317–328.
- André F. T. Martins and Ramón F. Astudillo. 2016. [From softmax to sparsemax: A sparse model of attention and multi-label classification](#). In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML’16, pages 1614–1623. JMLR.org.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’13, pages 3111–3119, USA. Curran Associates Inc.
- Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. [Key-value memory networks for directly reading documents](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409. Association for Computational Linguistics.
- Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. [Efficient and robust question answering from minimal context over documents](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 1725–1735. Association for Computational Linguistics.

Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2019. [Multi-style generative reading comprehension](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2273–2284, Florence, Italy. Association for Computational Linguistics.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. [Reasonet: Learning to stop reading in machine comprehension](#). In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17*, pages 1047–1055, New York, NY, USA. ACM.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018a. [Multi-granular sequence encoding via dilated compositional units for reading comprehension](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2141–2151, Brussels, Belgium. Association for Computational Linguistics.

Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. 2018b. [Densely connected attention propagation for reading comprehension](#). In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18*, pages 4911–4922, USA. Curran Associates Inc.

Yi Tay, Shuohang Wang, Anh Tuan Luu, Jie Fu, Minh C. Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. 2019. [Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4922–4931, Florence, Italy. Association for Computational Linguistics.

Ellen M. Voorhees. 2001. [The trec question answering track](#). *Natural Language Engineering*, 7(4):361–378.

Hugo Zaragoza, Nick Craswell, Michael J Taylor, Suchi Saria, and Stephen E Robertson. 2004. Microsoft cambridge at trec 13: Web and hard tracks. In *TREC*, volume 4, pages 1–1.