BookQA: Stories of Challenges and Opportunities

Stefanos Angelidis[1*], Lea Frermann[2], Diego Marcheggiani[3], Roi Blanco[3], Lluís Màrquez[3]

[1]Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh
[2]School of Computing and Information Systems, The University of Melbourne
[3]Amazon Research

In a nutshell

Book Question Answering (BookQA) is largely unexplored compared to other QA settings. Unique characteristics of books (length, literary language, lack of KBs, little training data) prohibit application of state-of-the-art QA methods.

Our contributions:

- We look at NarrativeQA's [1] *'Who'* questions which have book characters as answers.
- We evaluate a framework for predicting the correct answer from the full text of the book.
- We utilize pretraining on artificial questions.
- We discuss challenges of full-text BookQA and identify opportunities for improvements.

The Data:

NarrativeQA questions vary wildly in style. By constructing a corpus of 'Who' questions, we:

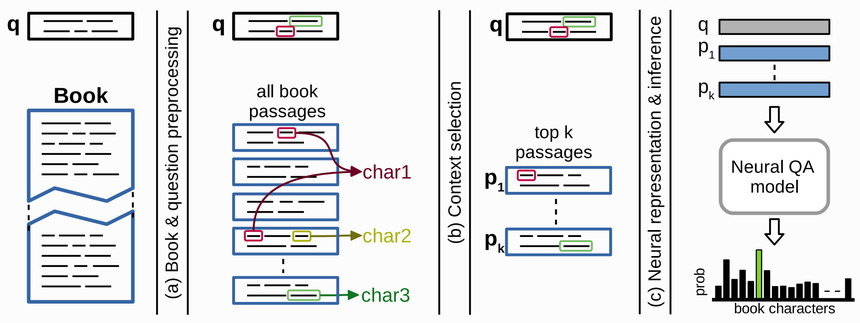
- Simplify output & evaluation (classification).
- Retain reasoning complexity of original.

Examples from corpus:

Who is Emily in love with? easier ↔ harder
Who is Emily imprisoned by?
Who helps Emily escape from the castle?
Who owns the castle in which Emily is imprisoned?
Who became Emily's guardian after her father's death?

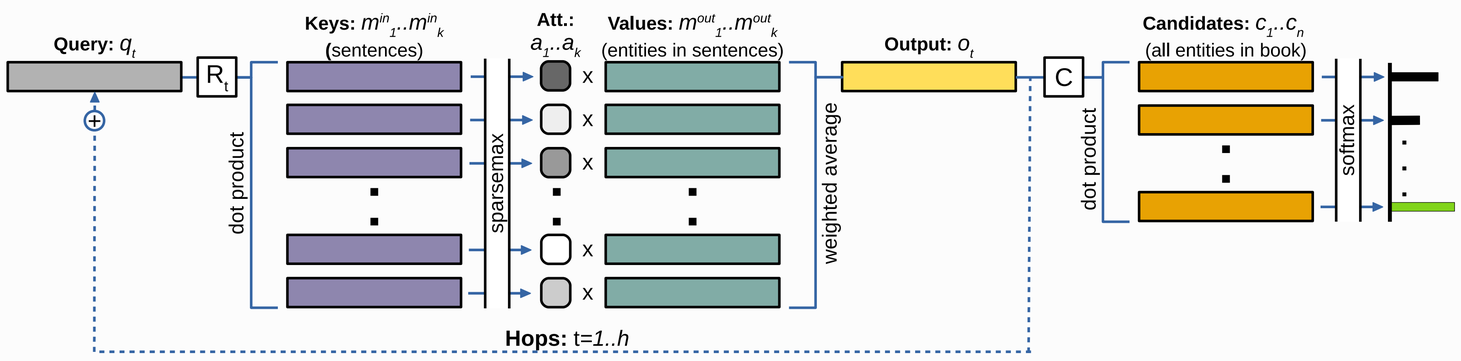
■ 3427 QA pairs from 614 books.

The Framework:



- **Preprocessing:** book-nlp parser [2]
- **Context Selection:** i) BM25F; ii) BERT-based
- **Neural Inference:** Variant of Key-Value MemNet [3]

Neural Inference with Key-Value Memory Network



Initialization:

$$\text{Query: } q_{t=0} = avg(v_{qw_1}, \ldots, v_{qw_m})$$
$$\text{Keys: } m^{in}_i = avg(v_{sw_1}, \ldots, v_{sw_l})$$
$$\text{Values: } m^{out}_i = avg(v_{c_1 \in s}, \ldots)$$
$$\text{Candidates: } c_j = v_{c_j}$$

At Hop t:

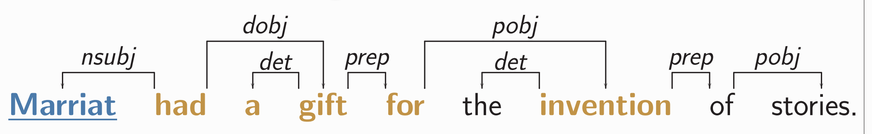
$$a_{ti} = sparsemax(q_t R_t m_i)$$
$$o_t = \sum_i a_{ti} m^{out}_i$$
$$q_{t+1} = q_t + o_t$$

After last hop:

$$p(c_j) = softmax(o_h C v_{c_j})$$

Pretraining with Artificial Questions

Original Sentence:



Artificial Question:

Who had a gift for invention?

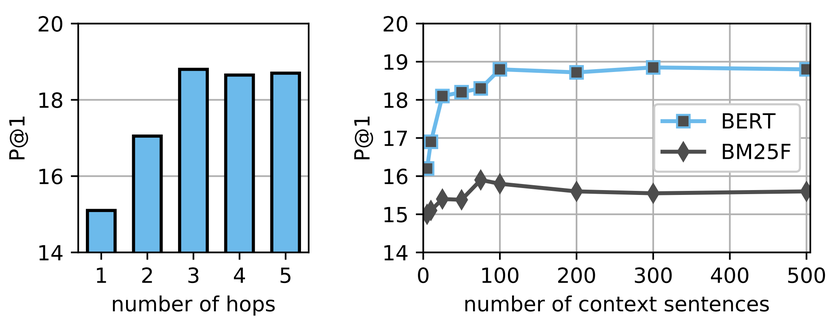
- **Problem:** not enough data for training inference model.
- **Our solution:** pretrain on artificially generated questions.
- We use source sentences where a book character is the subject or object of a verb.
- Simple rules and pruning over dependency tree.

Experimental Setup and Main Results

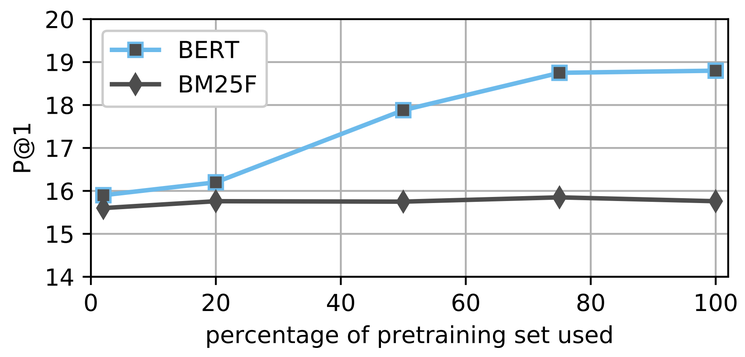
- **Context Selection:**
  - Top 20 passages (100 sentences).
- **Pretraining:**
  - Pretrain on artificial questions, using 20 previous sentences as context.
  - Fine tune on real questions.
- **Baselines:**
  - Most frequent character in book.
  - Most frequent character in context.

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
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

Further Results – Neural Inference



Further Results – Effect of Pretraining



Analysis – Question Answerability

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

- Mentions counted via book-nlp's character recognizer.
- Evidence identified via Amazon Mechanical Turk study.

Challenges & Opportunities

- **Inaccurate context selection:**
  - Book-tailored passage relevance
- **Vagueness of literary language:**
  - Paraphrase detection
  - Coreference resolution
  - Commonsense knowledge
- **Inadequate pretraining:**
  - Artificial questions that better resemble real ones (or other auxiliary task).

References

[1] Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, Edward Grefenstette. *The NarrativeQA Reading Comprehension Challenge.* (TACL 2018).

[2] David Bamman, Ted Underwood, Noah A. Smith. *A Bayesian Mixed Effects Model of Literary Character.* (ACL 2014).

[3] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, Jason Weston. *Key-Value Memory Networks for Directly Reading Documents.* (EMNLP 2016).

 *Corresponding author: s.angelidis@ed.ac.uk