Expand, Highlight, Generate: RL-driven Document Generation for Passage Reranking

Universiteit Leiden
Arian Askari, Mohammad Aliannejadi, Chuan Meng, Evangelos Kanoulas, Suzan Verberne
Leiden University and University of Amsterdam

EMNLP 2023

Study

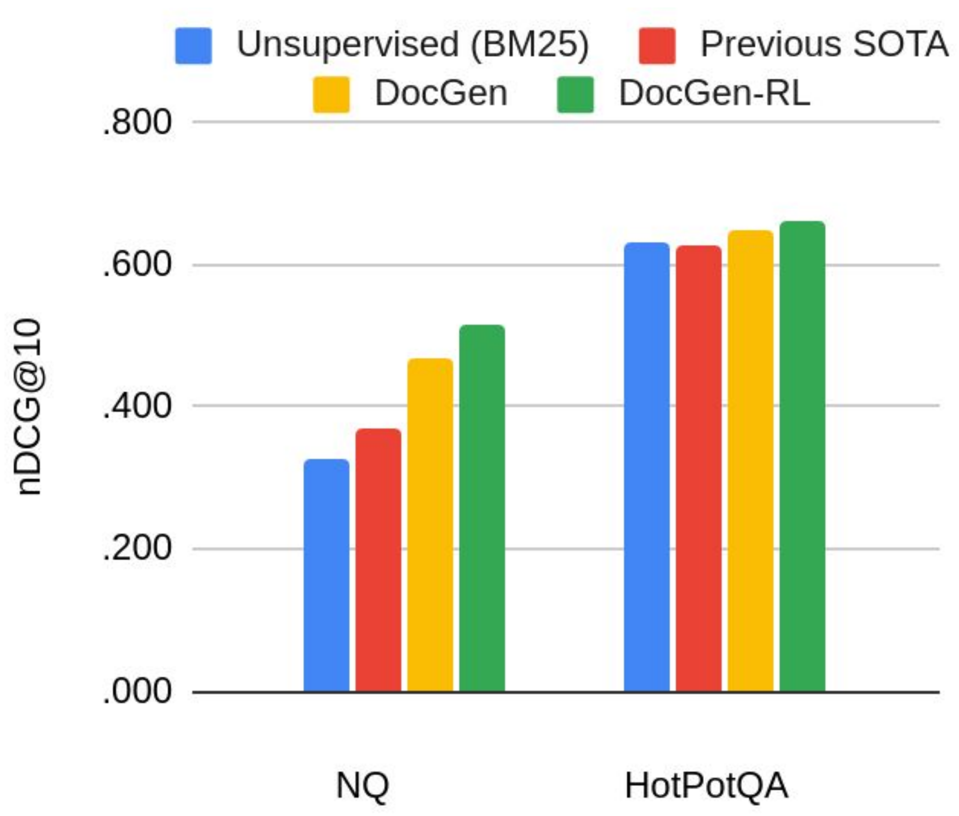
- We **address** augmenting training data for passage reranking models using **LLMs** in a novel direction:
  - Generating synthetic documents instead of synthetic queries
  - Figure 1 demonstrates the process
- We **propose** two novel synthetic document generators: **DocGen**, and **DocGen-RL.**
- **Advantages** for commercial search engines:
  - For **complex queries,** the relevant information may be spread across multiple documents.
  - For **new or trending queries,** optimizing retriever based on the recently received and repeated (trending) queries.

Proposed Methods: DocGen and DocGen-RL

- **DocGen** consists of a three-step pipeline (Figure 2) with few-shot learning:
  - **Query expansion**
  - **Query highlighting** using square brackets
  - **Document generation**, producing likely relevant documents given the expanded and highlighted query
- **DocGen-RL** uses reinforcement learning (RL) to optimize highlighting in the DocGen's pipeline with the estimated relevance of the document as a reward.

Main Results

- Both DocGen and DocGen-RL consistently outperform BM25 and previous state-of-the-art data augmentation methods.



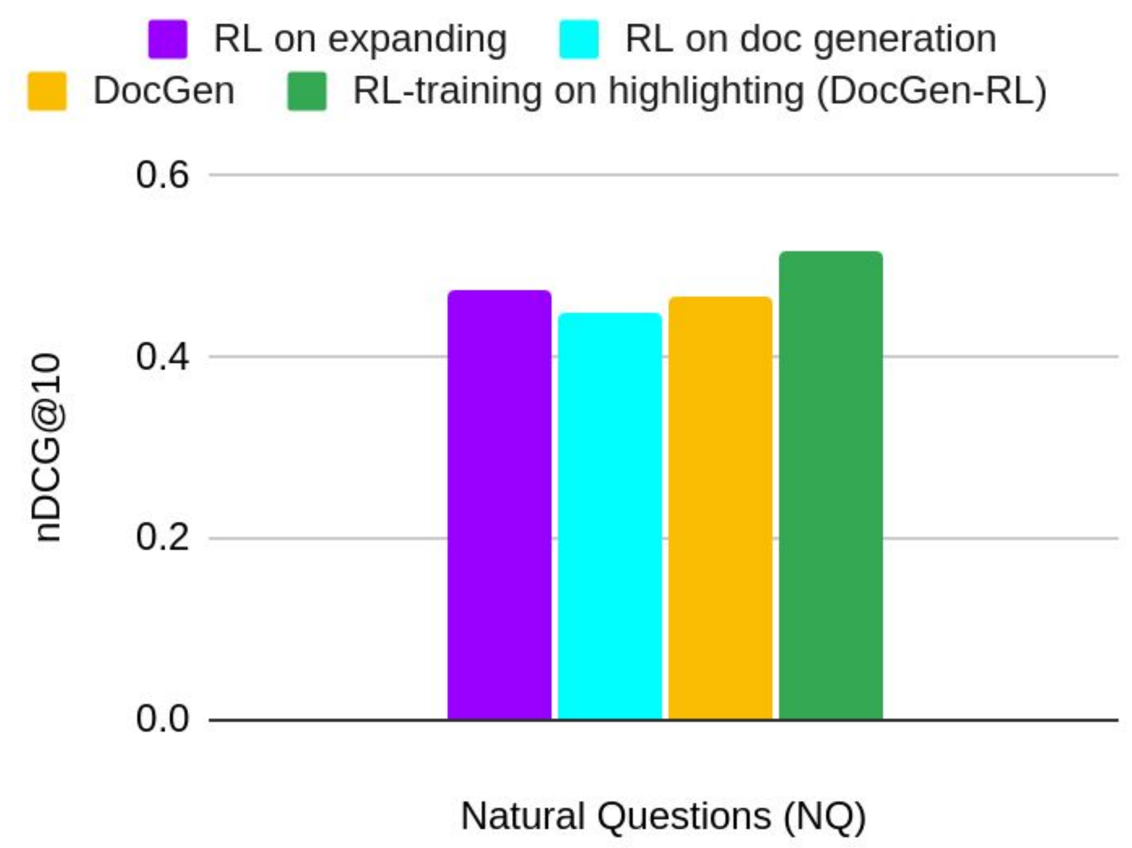
Scaling impact

- Significant improvement over the results by increasing the scale of either BLOOM or MonoT5 parameters.

Dataset	NQ-test
BLOOM 560M and T5-base (220M)	.467
BLOOM-3B	.482
T5-large (770M)	.495

RL-training analysis

- RL training for highlighting queries (DocGen-RL) achieves highest effectiveness.



Gap between synthetic and realistic data

- Similarity the human data in terms of the word overlap between query and document.

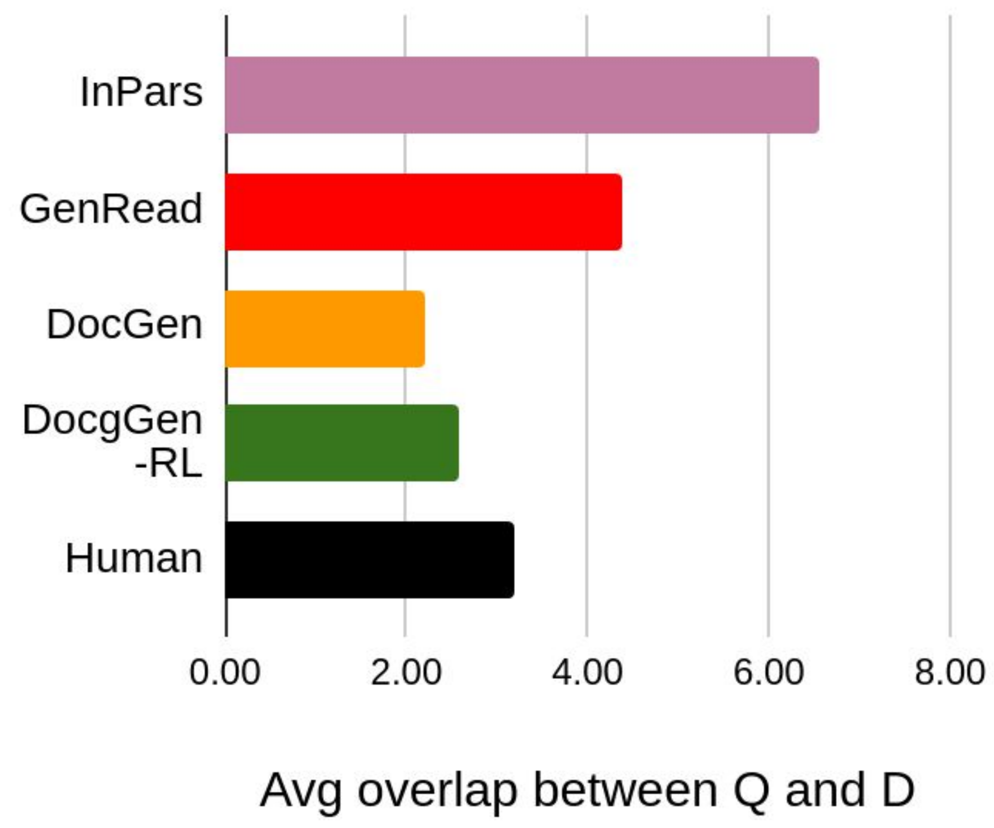


Figure 1. Illustration of training passage reranking using the DocGen.

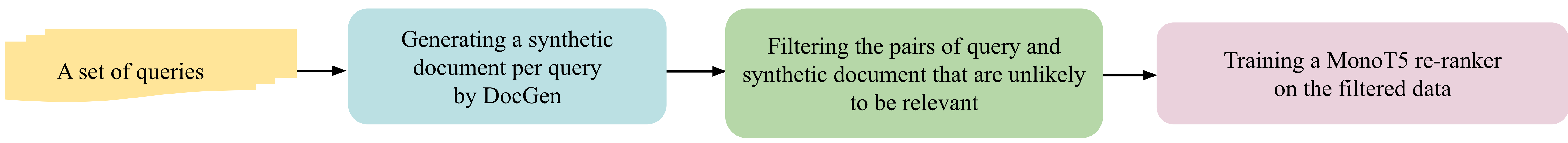


Figure 2. DocGen's pipeline