Autoregressive Search Engines: Generating Substrings as Document Identifiers

Dhyey Bhimani, Nisarg Ganatra, Harsh Raval and Shubham Shah

Abstract

Knowledge-intensive language tasks require NLP systems to provide correct answers and retrieve supporting evidence. While previous methods, like partitioning search spaces with hierarchical structures, have limitations, we propose using all n-grams in a passage as unique identifiers. This approach allows autoregressive models to efficiently generate and map n-grams to full passages, outperforming prior methods and achieving state-of-the-art results on the KILT benchmark with lower memory requirements.

1. Problem

Traditional retrieval systems often retrieve entire documents, missing the specific passages most relevant to a query, which leads to inefficiencies in information retrieval. Additionally, these systems impose rigid structures, such as hierarchical identifiers, or rely solely on query expansion, limiting their flexibility and effectiveness. They fail to fully leverage the advanced capabilities of modern autoregressive language models, such as context sensitivity, word order awareness, and probabilistic modeling. This gap affects their ability to precisely identify and rank relevant passages within large corpora. To overcome these limitations, there is a need for a retrieval system that integrates autoregressive models to generate n-grams, enabling accurate passage-level retrieval with enhanced efficiency and flexibility [1].

2. Dataset

The paper primarily utilizes datasets from Natural Questions (NQ) and the KILT benchmark to evaluate the proposed retrieval system. The NQ dataset consists of query-document pairs where each query is a question and the document is a Wikipedia page containing the answer span. The retrieval corpus for NQ is chunked into approximately 21 million passages, each containing 100 tokens, enabling passage-level retrieval. The KILT benchmark [2] integrates diverse datasets, including question answering, fact-checking, dialogue, and entity linking, all solvable by retrieving evidence from a unified Wikipedia corpus. For KILT, the corpus is re-chunked into around 36 million passages of 100 tokens, focusing on passage-level retrieval. Performance on these datasets is measured using metrics such as accuracy@k for NQ and R-precision for KILT, emphasizing retrieval accuracy at both passage and document levels. These datasets provide a comprehensive evaluation framework for assessing retrieval systems in knowledge-intensive tasks.

3. Evaluation Metric

The evaluation metrics used are Accuracy@k and R-Precision. Accuracy@k measures the fraction of queries for which at least one of the top-k retrieved passages contains the correct answer, commonly used for the Natural Questions (NQ) dataset. R-Precision, used for the KILT benchmark, focuses on precision-oriented retrieval by considering only the gold documents as correct answers. The formula for R-Precision is:

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

202311004 (D. Bhimani); 202311018 (N. Ganatra); 202311028 (H. Raval); 202311049 (S. Shah)

$$R\text{-Precision} = \frac{\text{Number of relevant passages retrieved}}{\text{Total number of relevant passages in the corpus}}$$

These metrics emphasize both the precision and relevance of the retrieved results, ensuring effective evaluation of retrieval performance.

4. Results

The results highlight the effectiveness of SEAL model over traditional methods like BM25 for passage-level retrieval tasks. While BM25 serves as a strong baseline, the SEAL model demonstrate superior performance due to their ability to leverage autoregressive modeling for n-gram generation and efficient passage mapping.

Model	FEV	T-REx	zsRE	NQ	HoPo	TQA	AVG
BM25	40.1	51.6	53.0	14.2	38.4	16.2	35.3
SEAL (LM+FM)	31.5	42.0	34.0	21.7	24.7	21.4	29.2
SEAL (LM+FM, intersective)	67.8	58.9	78.8	43.6	54.3	41.8	57.5
BM25 (Our)	30.1	42.6	46.0	10.4	40.2	14.2	30.5
SEAL (Our)	62.8	52.9	72.8	40.4	48.3	37.4	52.4

Table 1
Retrieval results on KILT with different configurations.

The improvements observed in our experiments suggest that proper tuning and configurations can significantly enhance the retrieval capabilities of these models. Additionally, the results underline the scalability and flexibility of SEAL's intersective approach, making it a robust choice for information retrieval tasks. These findings reinforce the value of modern autoregressive models in achieving state-of-the-art retrieval accuracy with efficiency.

5. Key Challenges and Learning

During the course of this project, we faced several challenges. A major challenge was the computational requirements, as the models we worked with were resource-intensive and required significant processing power for training. Additionally, managing the complexity of n-gram generation and ensuring efficient mapping to passages posed difficulties that required careful troubleshooting. We also encountered challenges in maintaining consistency and reproducibility of results across different datasets, which involved extensive debugging and fine-tuning.

On the learning side, this project provided valuable insights into the mechanisms of autoregressive retrieval systems and the role of n-grams in improving passage-level retrieval. We gained a deeper understanding of query optimization techniques and the principles behind efficient passage ranking. This experience has enhanced our ability to design and implement scalable solutions for information retrieval tasks while addressing practical challenges in the process.

References

[1] M. Bevilacqua, G. Ottaviano, P. Lewis, W. tau Yih, S. Riedel, F. Petroni, Autoregressive search engines: Generating substrings as document identifiers, 2022. URL: https://arxiv.org/abs/2204.10628. arXiv:2204.10628.

[2] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. D. Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel, S. Riedel, Kilt: a benchmark for knowledge intensive language tasks, 2021. URL: https://arxiv.org/abs/2009.02252. arXiv:2009.02252.