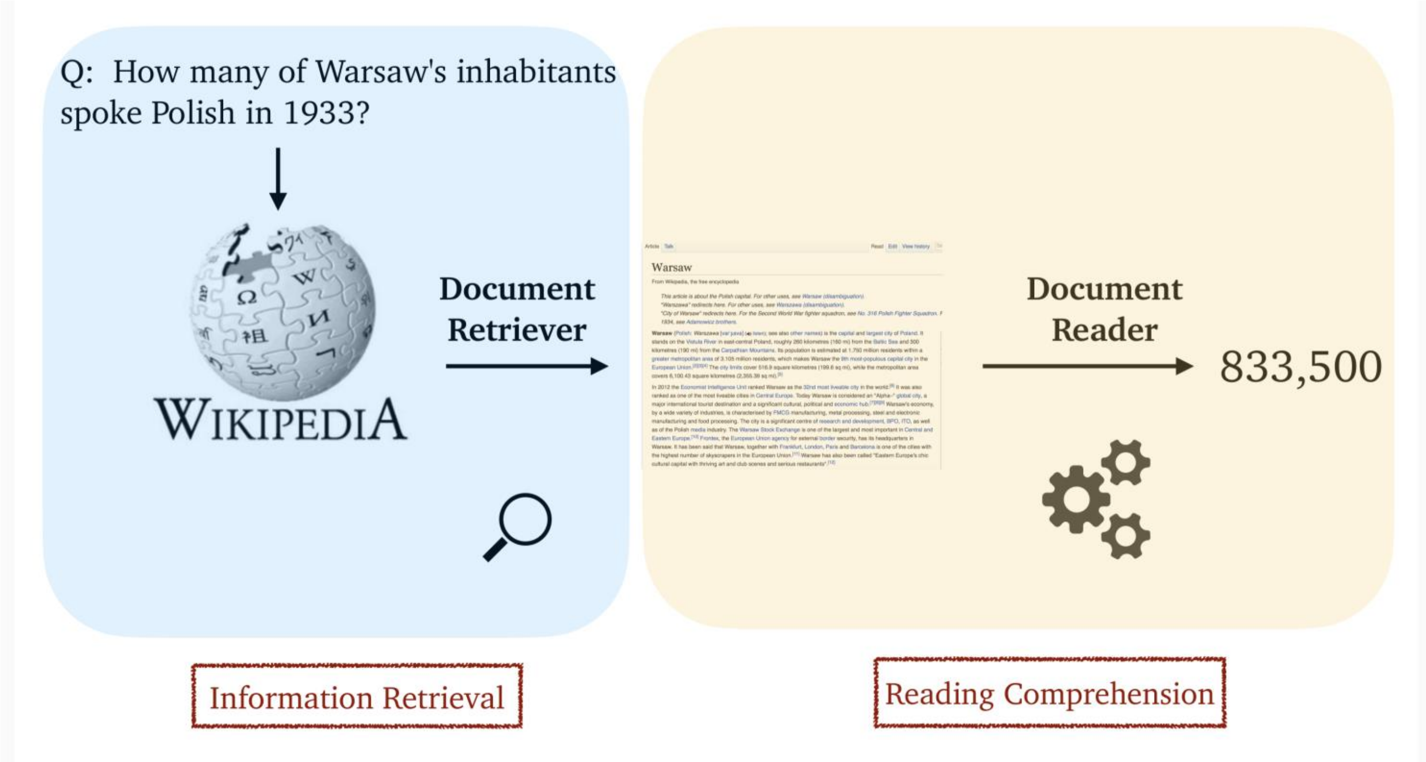
# Autoregressive Search Engines:
## Generating Substrings as Document Identifiers

Authors: Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen-tau Yih, Sebastian Riedel, Fabio Petroni

# Retriever-Reader Approaches

Using whole WIKIPEDIA (~5millions documents) as external memory.



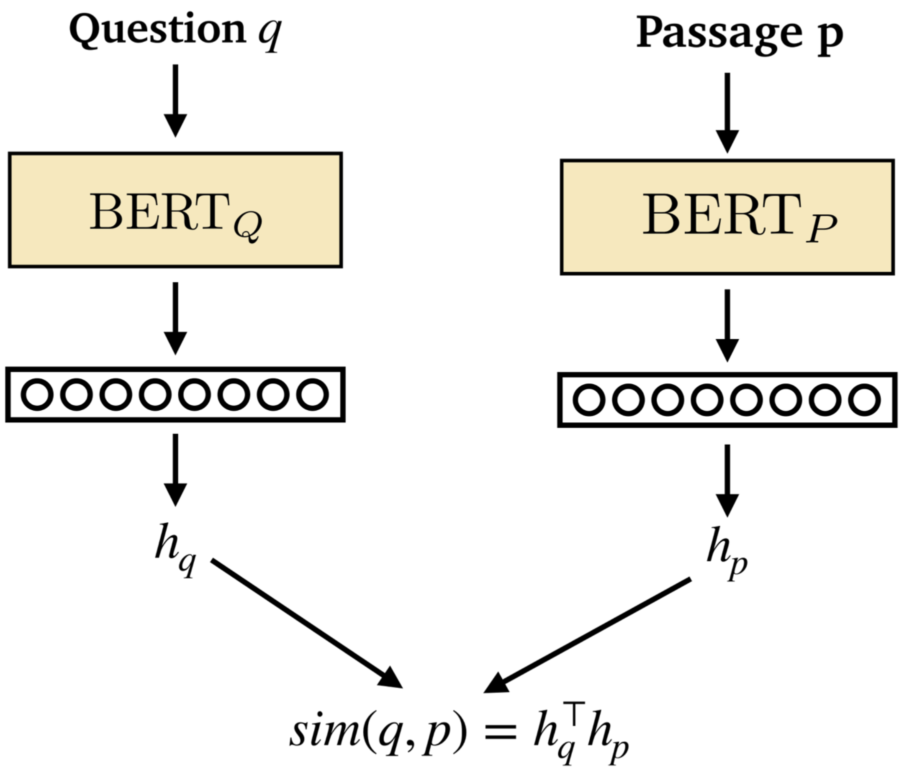

Cast as a reading comprehension problem,
Input a passage and a question.
Output is an Answer.

Reading Wikipedia to Answer Open-domain Questions,Chen et al., 2017.

[Borrowed from Ting's slides]

# Dense Passage Retrieval (DPR)

Directly training retriever with positive and negative passages.



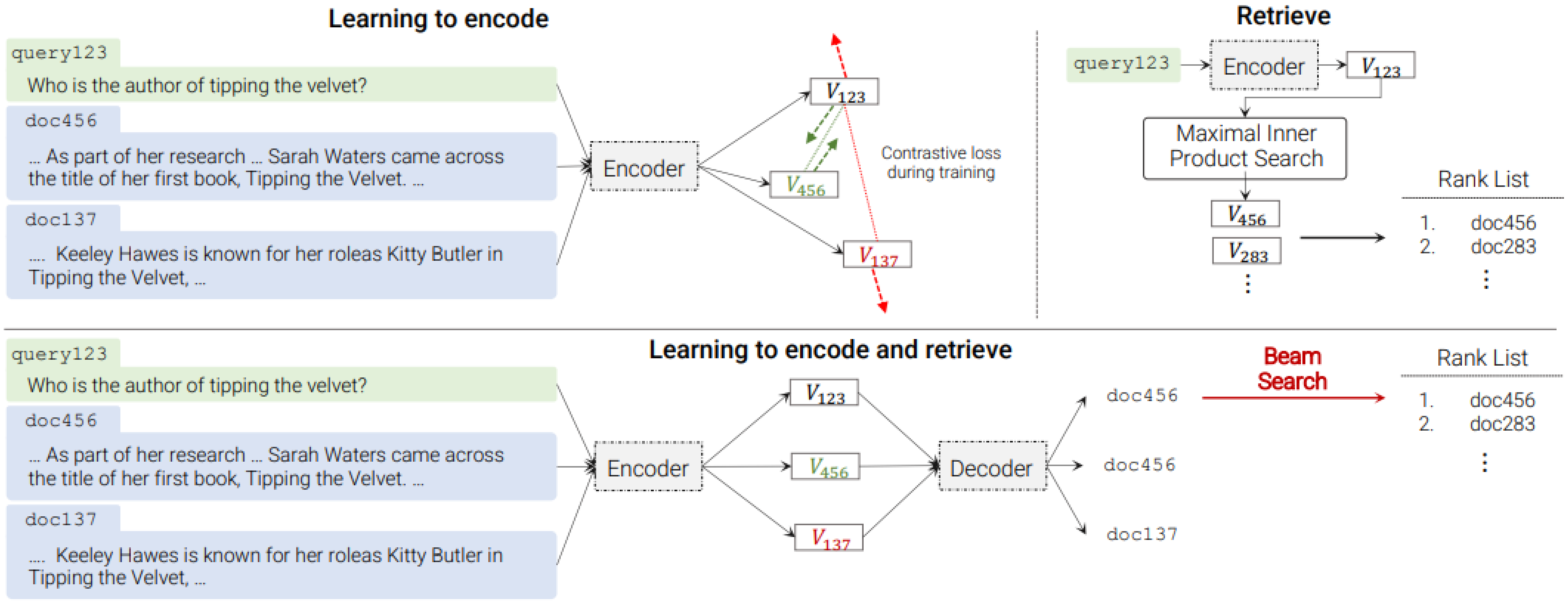

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^m$$

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-)$$
$$= -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\mathrm{sim}(q_i, p_{i,j}^-)}}.$$

Dense Passage Retrieval for Open-Domain Question Answering. Karpukhin et al., 2020

[Borrowed from Ting's slides]

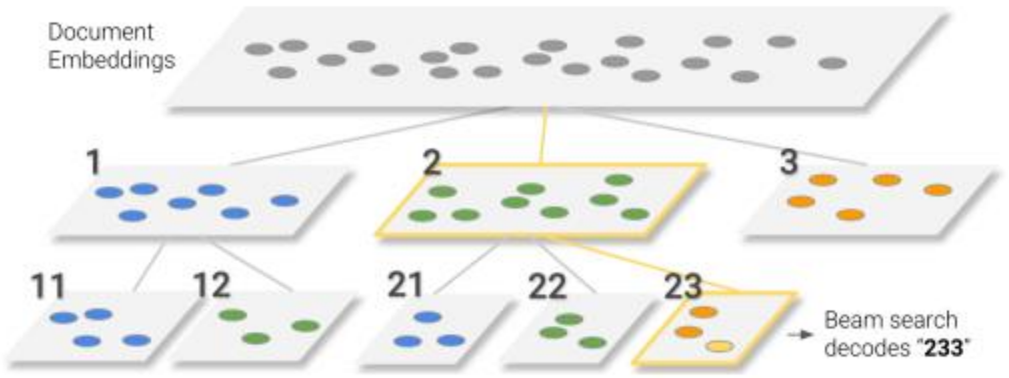
# Differentiable Search Index (DSI)




Transformer Memory as a Differentiable Search Index. Tay et al, 2022

# Differentiable Search Index (DSI)

## Semantically Structured Identifiers




---

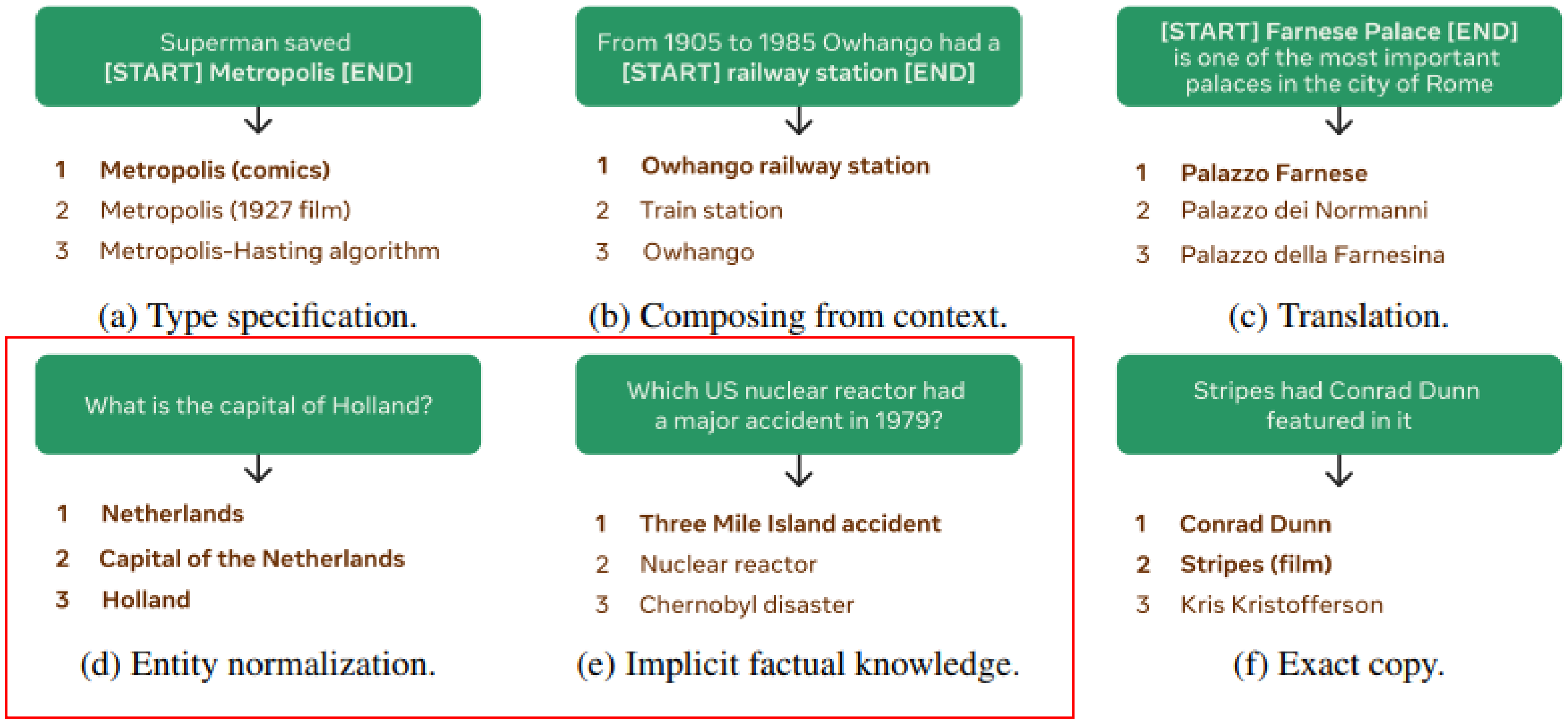
**Algorithm 1** Generating semantically structured identifiers

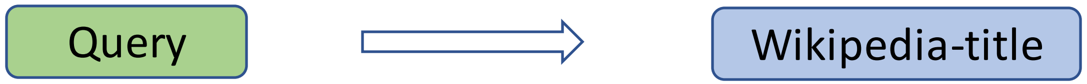
---

**Input:** Document embeddings $X_{1:N}$, where $X_i \in \mathbb{R}^d$
**Output:** Corresponding docid strings $J_{1:N}$
**function** GENERATESEMANTICIDS($X_{1:N}$)
  $C_{1:10} \leftarrow Cluster(X_{1:N},\ k = 10)$
  $J \leftarrow$ empty list
  **for** $i = 0$ **to** $9$ **do**
    $J_{current} \leftarrow [i] * |C_{i+1}|$
    **if** $|C_{i+1}| > c$ **then**
      $J_{rest} \leftarrow$GENERATESEMANTICIDS($C_{i+1}$)
    **else**
      $J_{rest} \leftarrow [0, \ldots, |C_{i+1}| - 1]$
    **end if**
    $J_{cluster} \leftarrow$elementwiseStrConcat($J_{current},\ J_{rest}$)
    $J \leftarrow J$.appendElements($J_{cluster}$)
  **end for**
  $J \leftarrow$ reorderToOriginal($J,\ X_{1:N},\ C_{1:10}$)
  **return** $J$
**end function**

---

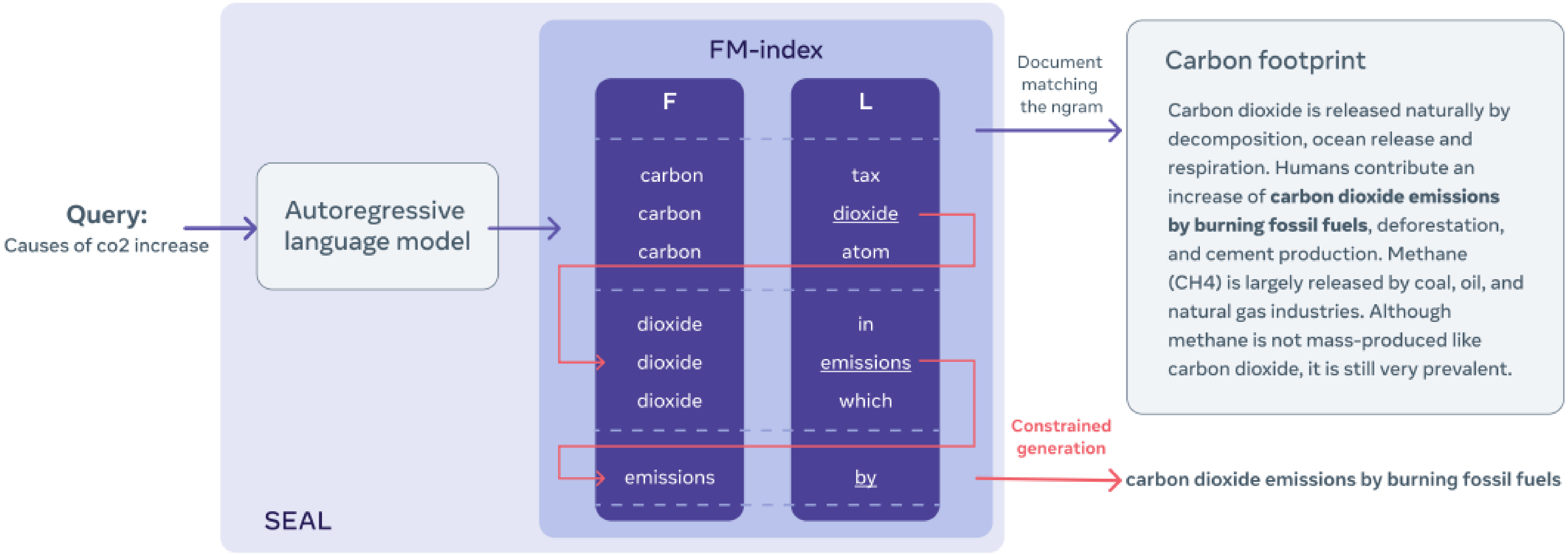
# Generative **En**tity **Re**trieval GENRE




[Autoregressive Entity Retrieval](#) De Cao et al., 2021

# Search Engines with Autoregressive LMs (SEAL)



# The FM-Index (Ferragina and Manzini. 2000)

- The FMindex can be used to count the frequency of any sequence of tokens n in O(|n|log|V|)
- For constrained decoding, the list of possible token successors can be obtained in O(|V|log|V|).
- the FM-index relies on the Burrows-Wheeler Transform (Burrows and Wheeler, 1994), or BWT

## Example: CABAC

| **F** | | | | | **L** |
|---|---|---|---|---|---|
| $\$^6$ | $C$ | $A$ | $B$ | $A$ | $C^5$ |
| $A^2$ | $B$ | $A$ | $C$ | $\$$ | $C^1$ |
| $A^4$ | $C$ | $\$$ | $C$ | $A$ | $B^3$ |
| $B^3$ | $A$ | $C$ | $\$$ | $C$ | $A^2$ |
| $C^5$ | $\$$ | $C$ | $A$ | $B$ | $A^4$ |
| $C^1$ | $A$ | $B$ | $A$ | $C$ | $\$^6$ |

- The first (F) and last (L) columns are the only ones that will be explicitly stored in the FM-index;
- we can locate any string <$\sigma_1, \sigma_2, \ldots, \sigma_n$> in the index by starting from $\sigma_n$ and going backwards
- FM-index implemented in `**sdsl-lite**`

Ferragina and Manzini. 2000. Opportunistic Data Structures with Applications https://people.unipmn.it/manzini/papers/focs00draft.pdf

# Autoregressive Retrieval (**LM** scoring)

- Constrained beam search with FM-Index
- Produce **fixed-length** *n*gram Candidates K (K = 10 in experiments)
- Each document is assigned the score (P(n|q)) of its most probable decoded occurring ngram

# Factoring in FM-index frequencies (**LM+FM** scoring)

$$P(n) = \frac{F(n, \mathcal{R})}{\sum_{d \in \mathcal{R}} |d|} \quad (1)$$

- P(n): ngram probability w.r.t. FM-index
- F(n, R): freq of ngram n in corpus R
- d: a doc in corpus R

$$w(n, q) = \max(0, \log \frac{P(n|q)(1 - P(n))}{P(n)(1 - P(n|q))}) \quad (2)$$

- P(n|q): ngram probability w.r.t. the seq2seq LM
- Smaller P(n) → ***distinctive*** *n*gram

# Intersective Scoring for Multiple Ngrams (**LM+FM intersective** scoring)

- A Problem with **LM** and **LM+FM** scoring
  - it is impossible to break ties among documents whose highest scoring ngram is the same, as they receive exactly the same score
- For doc d ∈ R, only consider K(d) ⊂ K
  - Remove overlapping ngrams with lower scores
- Scoring

$$W(d, q) = \sum_{n \in K^{(d)}} w(n, q)^{\alpha} \cdot \text{cover}(n, K^{(d)}) \quad (3)$$

the hyperparameters α, and β to, respectively, 2.0 and 0.8.

where $\alpha$ is a hyperparameter and the weight $\text{cover}(n, K)$ (controlled by the second hyperparameter $\beta$) is a function of how many ngram tokens are not included in the coverage set $C(n, K) \subset V$, *i.e.*, the union of all tokens in ngrams with a higher score. We define this coverage weight as follows:

$$\text{cover}(n, K) = 1 - \beta + \beta \cdot \frac{|\text{set}(n) \setminus C(n, K)|}{|\text{set}(n)|} \quad (4)$$

# Index Size

| System | Model Params | Size | Index Params | GPU? |
|---|---|---|---|---|
| *plain text* | - | 13.4GB | - | - |
| DPR | 220M | 64.6 GB | 16.1B | ✓ |
| BM25 | - | 18.8 GB | - | ✗ |
| GAR | 406M | 18.8 GB | - | ✗ |
| DSI-BART | 406M | - | - | - |
| SEAL | 406M | 8.8GB | - | ✗ |

Table 1: Language model and index size on Natural Questions (around 21M passages). SEAL's index is ~1.5 times smaller than uncompressed plain text.

# Results on NQ320K

| **System** | **hits@$k$** 1 | 10 |
|---|---|---|
| BM25 (`gensim`) | 15.3 | 44.5 |
| BM25 | 22.7 | 59.0 |
| DSI-BART | 25.0 | 63.6 |
| GENRE | **26.3** | 71.2 |
| SEAL (LM, $\lvert n \rvert = 3$) | 21.3 | 66.5 |
| SEAL (LM, $\lvert n \rvert = 4$) | 22.2 | 68.2 |
| SEAL (LM, $\lvert n \rvert = 5$) | 22.6 | 68.7 |
| SEAL (LM+FM) | 25.3 | 72.0 |
| SEAL (LM+FM, intersect.) | **26.3** | **74.5** |

Table 2: Results on NQ320$k$. Reporting hits@1 and hits@10. Best in bold.

# Results on NQ

| System | accuracy@$k$ 5 | 20 | 100 | Overlap? (A@100) ans. ✓ | ✗ | ques. ✓ | ✗ | EM |
|---|---|---|---|---|---|---|---|---|
| BM25 | 43.6 | 62.9 | 78.1 | 82.9 | 70.1 | 80.9 | 76.6 | 40.4 |
| DPR (Karpukhin et al., 2020) | **68.3** | **80.1** | 86.1 | 91.4 | 76.8 | 93.2 | 83.2 | 47.2 |
| GAR (Mao et al., 2021) | 59.3 | 73.9 | 85.0 | **91.6** | 74.4 | **94.1** | 80.4 | 46.2 |
| DSI-BART | 28.3 | 47.3 | 65.5 | 77.8 | 44.2 | 84.9 | 57.7 | 31.4 |
| Izacard and Grave (2021) | - | - | - | - | - | - | - | **48.2** |
| SEAL (LM, $\|n\| = 5$) | 40.5 | 60.2 | 73.1 | 82.2 | 57.1 | 85.2 | 64.9 | 36.0 |
| SEAL (LM+FM) | 43.9 | 65.8 | 81.1 | 86.9 | 70.9 | 89.5 | 78.1 | 42.9 |
| SEAL (LM+FM, intersective) | 61.3 | 76.2 | **86.3** | 91.2 | **77.7** | 93.2 | **84.1** | 48.0 |

# Results on KILT

| Model | FEV | T-REx | zsRE | NQ | HoPo | TQA | WoW | AVG |
|---|---|---|---|---|---|---|---|---|
| BM25 | 40.1 | 51.6 | 53.0 | 14.2 | 38.4 | 16.2 | 18.4 | 33.1 |
| DPR (Maillard et al., 2021) | 43.9 | 58.5 | **78.8** | 28.1 | 43.5 | 23.8 | 20.7 | 42.5 |
| MT-DPR (Maillard et al., 2021) | 52.1 | 53.5 | 41.7 | 28.8 | 38.4 | 34.2 | 24.1 | 39.0 |
| MT-DPR (Oğuz et al., 2021) | 52.1 | **61.4** | 54.1 | 40.1 | 41.0 | 34.2 | 24.6 | 43.9 |
| MT-DPR† (Oğuz et al., 2021) | 61.4 | 68.4 | 73.3 | 44.1 | 44.6 | 38.9 | 26.5 | 51.0 |
| MT-DPR† (large) (Oğuz et al., 2021) | 62.8 | 66.6 | 66.9 | 42.6 | 42.1 | 37.9 | 23.4 | 48.9 |
| SEAL (LM+FM) | 31.5 | 42.0 | 34.0 | 21.7 | 24.7 | 21.4 | 17.6 | 27.6 |
| SEAL (LM+FM, intersective) | **67.8** | 58.9 | **78.8** | **43.6** | **54.3** | **41.8** | **36.0** | **54.5** |

Table 4: Retrieval results on individual KILT dev set(s), with the average in the rightmost column. Reporting passage-level R-precision (higher is better). We mark model that are also trained on additional synthetic data (Lewis et al., 2021c) with †. All SEAL models are multitask. Best among models trained only on KILT queries in bold.

# Results on Downstream Tasks

| System | FEV<br>ACC | T-REx<br>ACC | zsRE<br>ACC | NQ<br>EM | HoPo<br>EM | TQA<br>EM | WoW<br>F1 |
|---|---|---|---|---|---|---|---|
| KGI (Glass et al., 2021)† | 85.6 | **84.4** | 72.6 | 45.2 | - | 61.0 | 18.6 |
| Hindsight (Paranjape et al., 2021) | - | - | - | - | - | - | **19.2** |
| DPR+BART (Petroni et al., 2021) | 86.7 | 59.2 | 30.4 | 41.3 | 25.2 | 58.6 | 15.2 |
| RAG (Petroni et al., 2021) | 86.3 | 59.2 | 44.7 | 44.4 | 27.0 | 71.3 | 13.1 |
| MT-DPR+BART (Maillard et al., 2021) | 86.3 | - | 58.0 | 39.8 | 31.8 | 59.6 | 15.3 |
| MT-DPR+FiD (Piktus et al., 2021) | 89.0 | 82.5 | 71.7 | 49.9 | 36.9 | 71.0 | 15.7 |
| MT-DPR-WEB+FiD (Piktus et al., 2021) | 89.0 | 81.7 | 74.2 | 51.6 | 38.3 | **72.7** | 15.5 |
| SEAL+FiD (LM+FM) | 87.9 | 83.7 | 74.2 | 47.3 | 37.6 | 65.8 | 17.5 |
| SEAL+FiD (LM+FM, intersective) | **89.5** | 83.6 | **74.7** | **53.7** | **40.5** | 70.9 | 18.3 |

Table 5: Downstream results on the KILT test set(s). Downstream metrics are accuracy (FEVER, T-REx, zero-shot RE), exact match (Natural Questions, HotpotQA, TriviaQA), or F1 (Wizard of Wikipedia). Best in bold. †: result taken from the `eval.ai` KILT leaderboard.

# Ablation Study

| System | Constr. | Beam | A@20 | A@100 |
|---|---|---|---|---|
| SEAL (LM+FM) | ✓ | 15 | 65.8 | 81.1 |
| | ✗ | 15 | 65.3 | 80.1 |
| | ✓ | 3 | 63.3 | 78.0 |
| | ✓ | 5 | 64.7 | 79.9 |
| | ✓ | 10 | 65.4 | 80.8 |
| SEAL (LM+FM, intersective) | ✓ | 15 | 76.2 | 86.3 |
| | ✗ | 15 | 76.2 | 86.2 |
| | ✓ | 3 | 75.2 | 84.9 |
| | ✓ | 5 | 75.9 | 85.8 |
| | ✓ | 10 | 76.4 | 86.4 |

Table 6: Ablation on Natural Questions. SEAL when using (✓) or not using (✗) FM-index constrained decoding, for beam size values in $\{3, 5, 10, 15\}$. Reporting accuracy@$k$.

# Case Study

| score | # | identifier | doc #1 | doc #2 |
|---|---|---|---|---|
| 273.2 | 1 | earthquakes can be predicted | **Seismology** @@ for precise **earthquake predictions**, including the VAN method. Most **seism**ologists do not believe that a system to provide timely warnings for individual **earthquakes** has yet been developed, and many believe that such a system would be unlikely to give **useful** warning of impending **seismic** events. However, more general **forecasts** routinely **predict** seismic **hazard**. Such **forecasts estimate** the **probability** of an **earthquake** of a particular [...] | **Earthquake prediction** @@ reliably identified across significant spatial and temporal scales. While part of the scientific community hold that, taking into account non-**seismic** precursors and given enough resources to study them extensively, **prediction** might be **possible**, most scientists are pessimistic and some maintain that **earthquake prediction** is inherently impossible. **Predictions** are deemed significant if they can be shown to be successful beyond random chance.[...] |
| 272.7 | 75 | Earthquake prediction @@ | | |
| 269.9 | 3 | predicted earthquakes | | |
| 229.7 | 11 | Earthquake forecasting @@ | | |
| 217.2 | 2 | prediction Earthquake | | |
| 211.5 | 1 | used to predict earthquakes | | |
| 205.3 | 7 | earthquakes. Earthquake | | |
| | – | | | |
| –77.0 | 9 | Seismic metamaterial @@ | | |
| –97.4 | 14 | Seismic risk in Malta @@ | | |
| –113.4 | 3 | Quaternary (EP) @@ | | |
| –150.3 | 1 | used to predict the locatio[...] | | |
| –301.5 | 17 | Precipice (Battlestar Gala[...] | | |

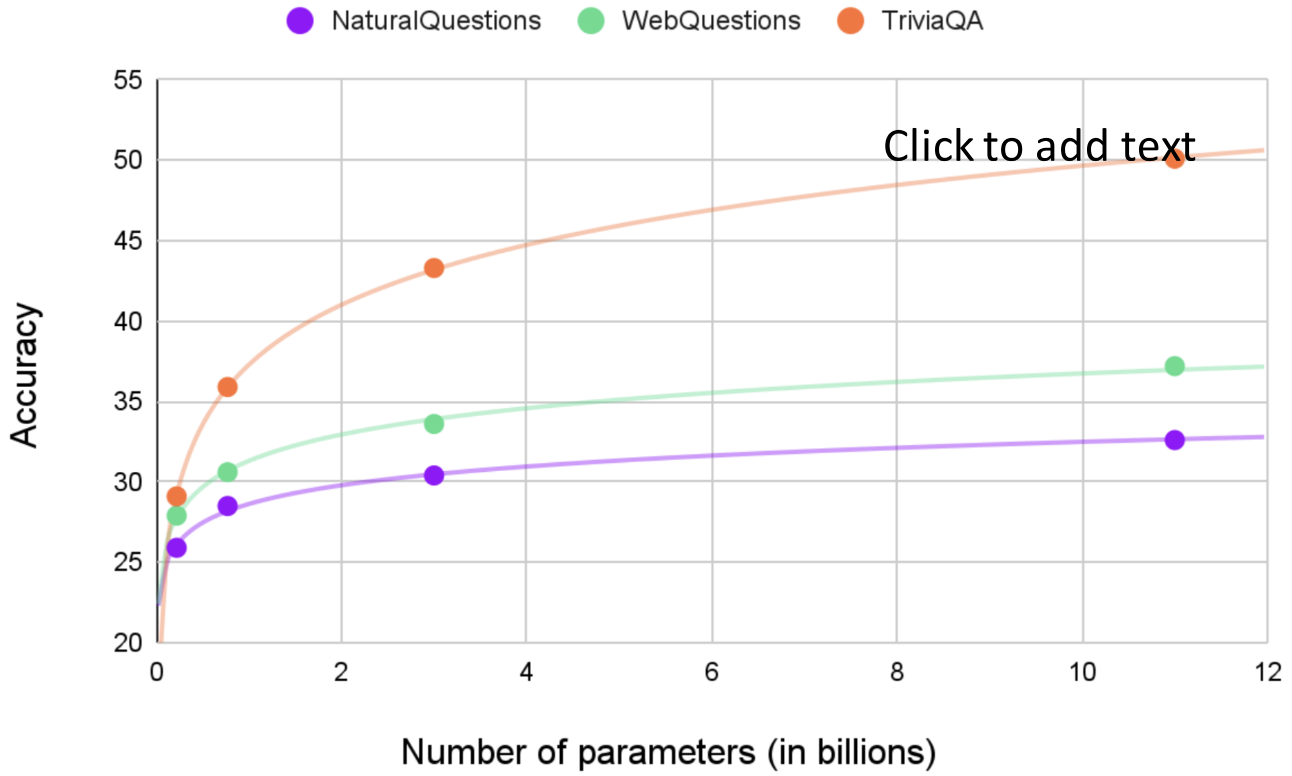
Table 7: Best (top) and worst (bottom) generated keys for the query “can you predict earthquakes” (left), and retrieved documents (right). Matched ngrams in bold. “@@” separates title and body.

# Open Discussions

- Next Steps:
  - Larger Model?
    - BART Large (400M) is very small today
    - PaLM 540B, 1000+ x larger
  - Better doc identifiers?
- Future of Search & QA

# Retriever Free Approaches

- Can we use pre-trained language models to act as “knowledge storage”?




The performance is largely impacted by the model size.

T5 on Three Open-domain QA Datasets

[Borrowed from Ting’s slides]

How Much Knowledge Can You Pack Into the Parameters of a Language Model? (Adam Roberts, Colin Raffel, Noam Shazeer)

Thanks