Introduction to Generative Retrieval

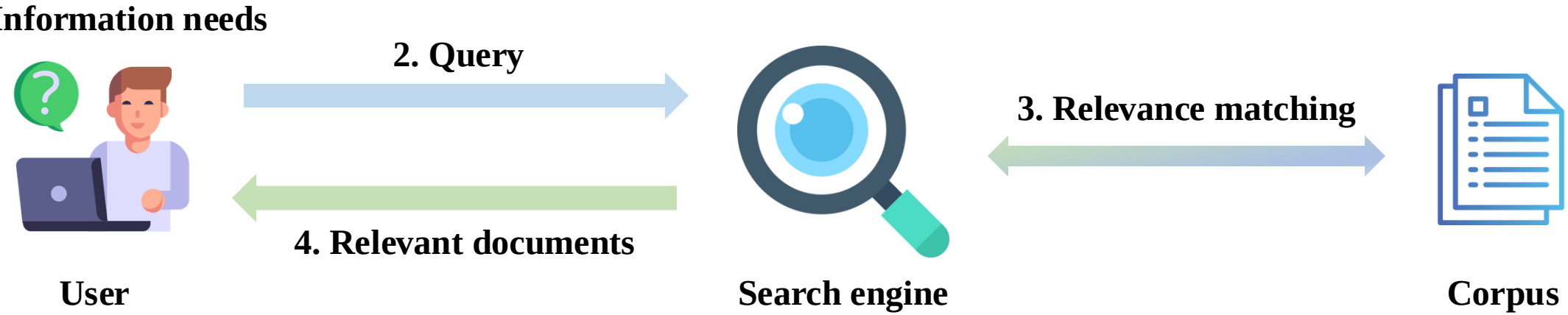
Yubao Tang (唐钰葆)
中国科学院计算技术研究所 博士三年级
tangyubao21b@ict.ac.cn

Content

- Backgrounds & preliminaries
- Semantic-Enhanced Differentiable Search Index Inspired by Learning Strategies
- Related work of our team

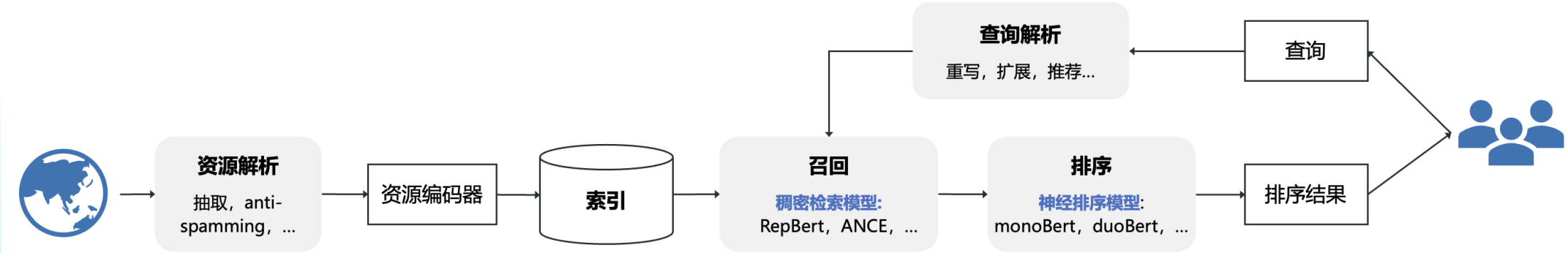
Backgrounds & preliminaries

Information Retrieval



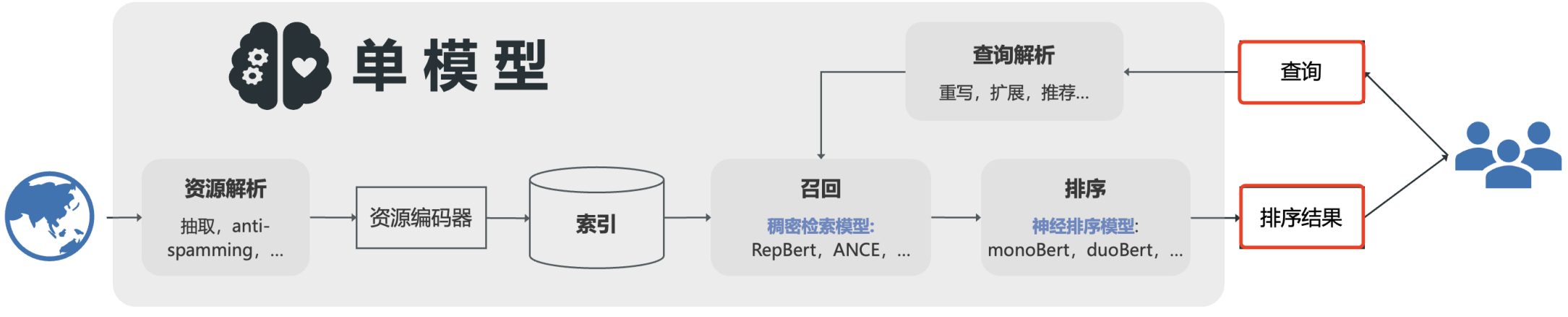
Information Retrieval

- **Index-retrieve-then-rank**
  - building the external index for the corpus
  - retrieving an initial set of candidate documents for a query
  - determining the relevance degree of each candidate



Information Retrieval

- **Limitations**
  - during training, heterogeneous modules with different optimization objectives may lead to sub-optimal performance, and capturing fine-grained relationships between queries and documents is challenging
  - during inference, a large document index is needed to search over the corpus, leading to substantial memory and computational requirements



Matching V.S. Association

Rethinking Search: Making Domain Experts out of Dilettantes, 2021, SIGIR Forum

Generative retrieval (GR) Differentiable Search Index

- Formalizing the document retrieval task as a Seq2Seq problem
- The information of all the documents within a corpus is encoded into the model parameters.
- Directly mapping string queries to relevant document identifiers (docids)

$$p(Relevant\ Results \mid Query)$$

- **Advantages**
  - Enabling the end-to-end optimization
  - Supporting fine-grained interaction with the model parameters
  - The autoregressive decoding significantly reduces the memory space and computational cost

Rethinking Search: Making Domain Experts out of Dilettantes, 2021, SIGIR Forum
Transformer Memory as a Differentiable Search Index, 2021, NeurIPS

Two basic tasks of GR

- **The indexing task： memorizing the corpus**
  - Learning a mapping from the document content to its identifier (docid)
  - The index is stored in model parameters, and indexing is simply another kind of model training

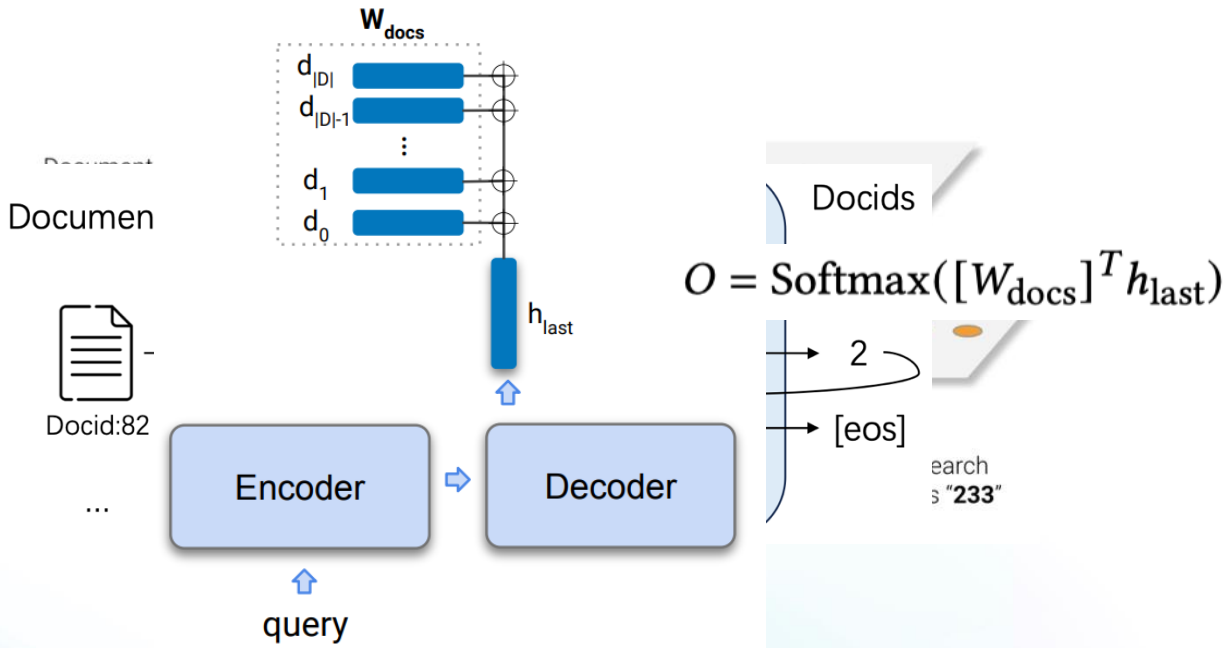
$$L_{indexing}(D, I_D; \theta) = -\sum_{d \in D} \log P(id|d; \theta)$$

- **The retrieval task: modelling the relevance**
  - Mapping queries to relevant docids

$$L_{retrieval}(Q, I_D^Q; \theta) = -\sum_{q \in Q} \log P(id^q|q; \theta)$$

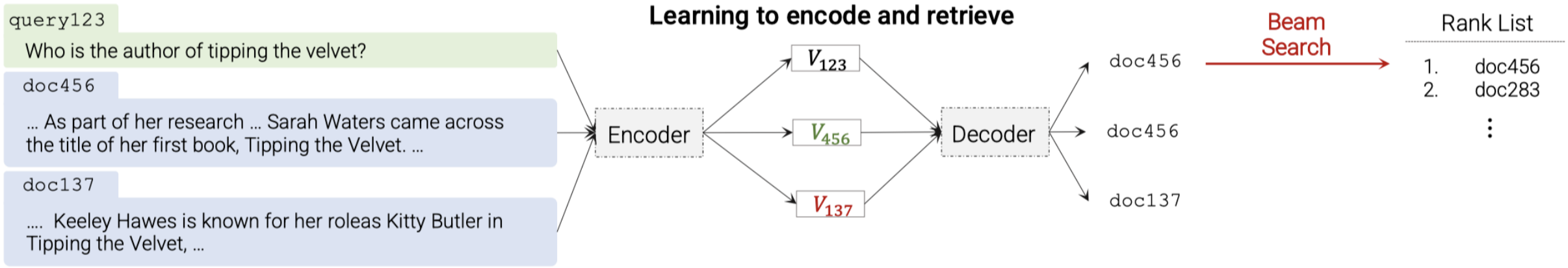
Docids of DSI

- Atomic Docid
- Naive String Docid
- Semantic String Docid



Transformer Memory as a Differentiable Search Index, 2021, NeurIPS

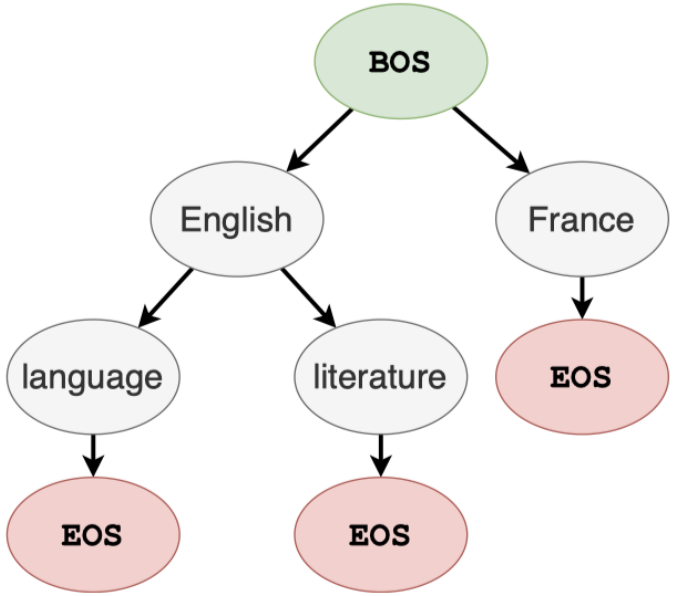
Basic optimization & inference



$$L_{total} = L_{indexing}(D, I_D; \theta) + L_{retrieval}(Q, I_D^Q; \theta)$$
$$= -\sum_{d \in D} \log P(id|d; \theta) - \sum_{q \in Q} \log P(id^q|q; \theta)$$

$$i_t = Model(q, i_0, \ldots, i_t - 1)$$

Inference: constrained beam search



Performance

Dataset	$\|D\|$	Train Pairs	Val Pairs	V_{doc_out}
NQ10K	10K	8K	2K	320K
NQ100K	86K	80K	20K	320K
NQ320K	228K	290K	17K	320K

Table 3: Experimental results on NQ document retrieval. DSI outperforms BM25 and Dual Encoder baselines. Among all the Docid representation methods, Semantic String Docids perform the best.

				NQ10K		NQ100K		NQ320K	
Model	Size	Params	Method	Hits@1	Hits@10	Hits@1	Hits@10	Hits@1	Hits@10
BM25	-	-	-	12.4	33.5	20.9	46.4	11.6	34.4
T5	Base	220M	Dual Encoder	16.2	48.6	18.7	55.2	20.5	58.3
T5	Large	800M	Dual Encoder	18.8	55.7	22.3	60.5	22.4	63.3
T5	XL	3B	Dual Encoder	20.8	59.6	23.3	63.2	23.9	65.8
T5	XXL	11B	Dual Encoder	22.1	61.6	24.1	64.5	24.3	67.3
DSI	Base	250M	Atomic Docid	13.0	38.4	23.8	58.6	20.7	40.9
DSI	Large	800M	Atomic Docid	31.3	59.4	17.1	52.3	6.9	27.3
DSI	XL	3B	Atomic Docid	40.1	76.9	19.0	55.3	28.1	61.9
DSI	XXL	11B	Atomic Docid	39.4	77.0	25.3	**67.9**	24.0	55.1
DSI	Base	250M	Naive String Docid	28.1	48.0	18.7	44.6	6.7	21.0
DSI	Large	800M	Naive String Docid	34.7	60.5	21.2	50.7	13.3	19.9
DSI	XL	3B	Naive String Docid	44.7	66.4	24.0	55.1	16.7	58.1
DSI	XXL	11B	Naive String Docid	46.7	**77.9**	**27.5**	62.4	23.8	55.9
DSI	Base	250M	Semantic String Docid	33.9	57.3	19.0	44.9	27.4	56.6
DSI	Large	800M	Semantic String Docid	37.5	65.1	20.4	50.2	35.6	62.6
DSI	XL	3B	Semantic String Docid	41.9	67.1	22.4	52.2	39.1	66.8
DSI	XXL	11B	Semantic String Docid	**48.5**	72.1	26.9	59.5	**40.4**	**70.3**

Transformer Memory as a Differentiable Search Index, 2021, NeurIPS

What aspects need optimization to improve the model's retrieval performance?

Key aspects

- **How to assign an identifier to each document ?**
- **How to memorize the corpus better ?**
- **How to model the relevance better ?**
- **……**

$$L_{total} = L_{indexing}(D, I_D; \theta) + L_{retrieval}(Q, I_D^Q; \theta)$$
$$= -\sum_{d \in D} \log P(id|d; \theta) - \sum_{q \in Q} \log P(id^q|q; \theta)$$

Existing methods

- **How to assign an identifier to each document ?**
    - Number-based:
        - Atomic Docid
        - Naive String Docid
        - Semantic String Docid
    - Word-based:
        - Titles
        - N-grams
- **How to memorize the corpus ?**
    - Taking the documents as inputs and generates docids as outputs

Semantic-Enhanced Differentiable Search Index Inspired by Learning Strategies

KDD 2023

Yubao Tang[1], Ruqing Zhang[1], Jiafeng Guo[1], Jiangui Chen[1],
Zuowei Zhu[2], Shuaiqiang Wang[2], Dawei Yin[2], Xueqi Cheng[1]

{tangyubao21b, zhangruqing, guojiafeng, chenjiangui18z, cxq}@ict.ac.cn
{zhuzuowei,wangshuaiqiang}@baidu.com
yindawei@acm.org

Memory and recall for humans

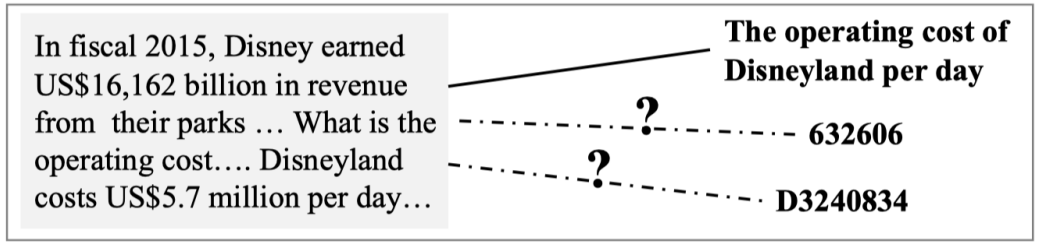
- **How does the human mind remember documents**
  - Learning Strategies in Cognitive Psychology are intended to influence learner's encoding process

In fiscal 2015, Disney earned US$16,162 billion in revenue from their parks … What is the operating cost…. Disneyland costs US$5.7 million per day…

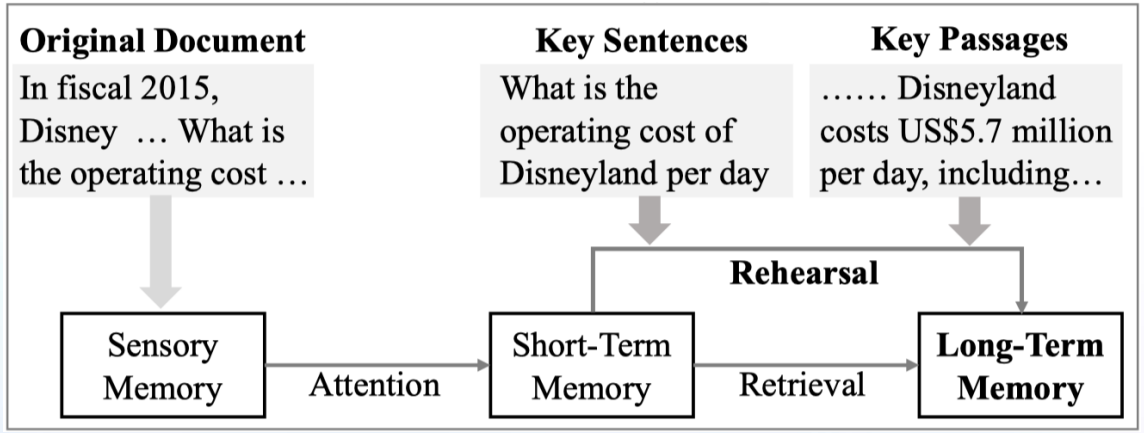
The teaching of learning strategies, 1983, Innovation abstracts

Memory and recall for humans

- **Elaboration Strategies**
  - Naming a document with natural language words which have semantic relationships with it, would contribute to better encoding and recall for human brain



- **Rehearsal Strategies**
  - Ones who underline the important contents in a document are able to recall substantially more information and higher long-term

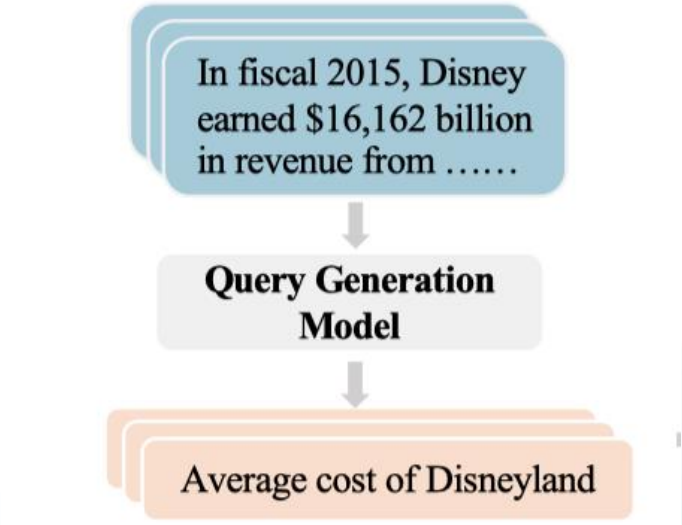


Cognition: Exploring the science of the mind, 1997, WW Norton & Co
Cognitive psychology, 2005, Pearson Education New Zealand

Approach

Approach

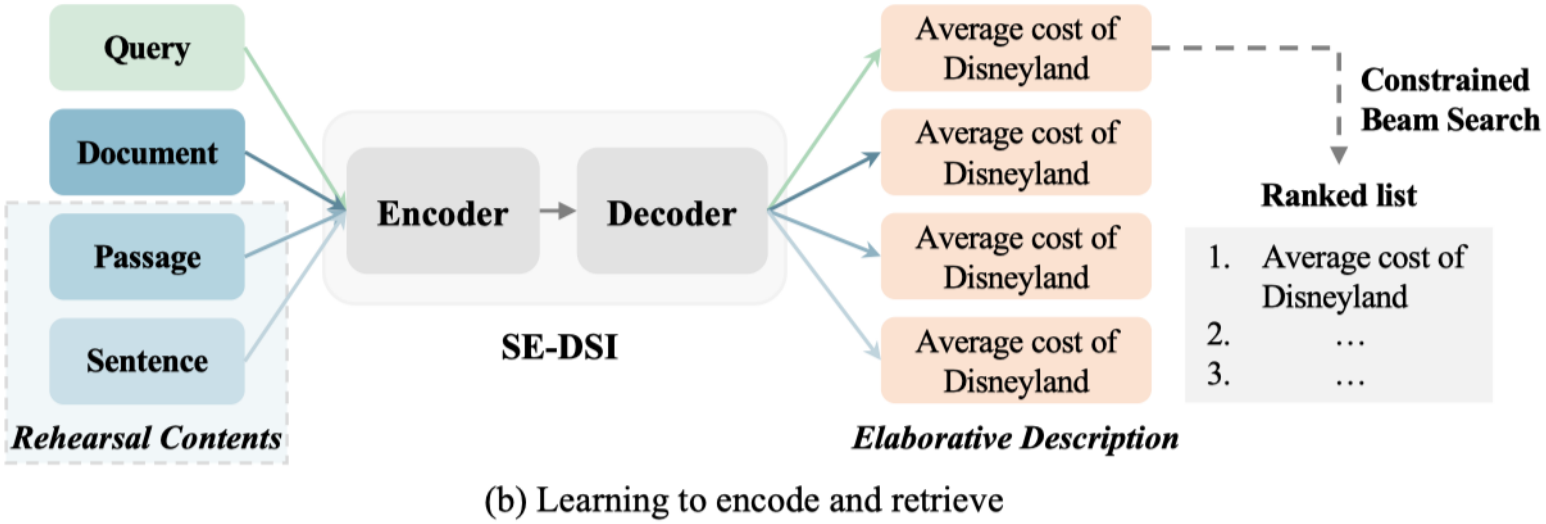
- **A novel Semantic-Enhanced DSI model (SE-DSI)**
- **Elaborative Description(ED)**:
  - Describing the identifiers in natural language.
  - We propose to generate ED for each document by off-the-shelf DocT5query Model.



(a) Elaborative Description generation

Approach

- **Rehearsal Contents(RCs)**:
  - We propose to select multiple important parts in a document as RCs, and the original document augmented with RCs are used to memorize the original document.



(b) Learning to encode and retrieve

Approach

- **Rehearsal Contents(RCs)**

Informative

Fluency

Diversity

Approach

- **Rehearsal Contents(RCs)**
  - **Leading-style**
    - directly use the leading passages and sentences of each original document as RCs
  - **Summarization-style**
    - the important information from the local context (e.g., sentence-level) and the broader context (e.g., paragraph-level).
    - a part is important in a document if it is highly related to many important parts

Training & Inference

- **Training**

$$\mathcal{L}(\theta) = \sum_{d_i \in \mathcal{D}} logP(ED_i|SE_\theta(d_i)) + \sum_{d_i \in \mathcal{D}} logP(ED_i|SE_\theta(RC_i^p)) + \sum_{d_i \in \mathcal{D}} logP(ED_i|SE_\theta(RC_i^s)) + \sum_{q_j \in Q} logP(ED_i|SE_\theta(q_j)),$$

- **Inference**
  - constrained beam search

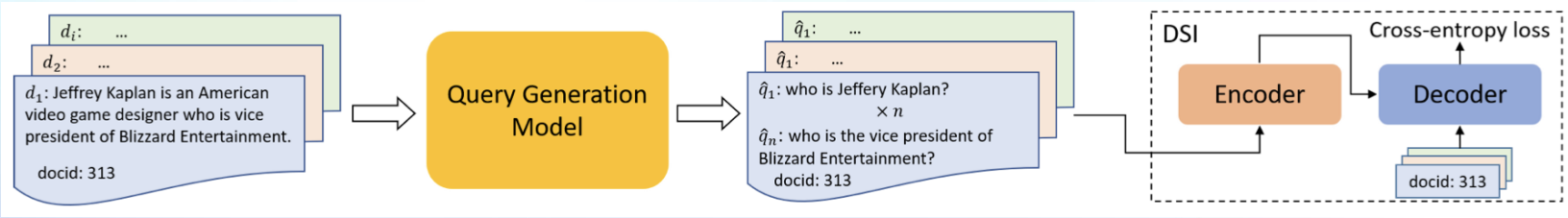
Offline and online experiments

Offline experiments settings

- **Datasets**

Dataset	#Doc	#Train	#Dev
MS MARCO 10K	13,569	14,763	1,330
MS MARCO 100K	89,154	96,948	3,000
MS MARCO Full	3,213,835	367,013	5,193
NQ 100K	100,000	100,853	2,800

- **Baselines**
  - **Traditional document retrieval methods**
    - BM25, RepBERT
  - **GR methods**
    - DSI-ARB(arbitrary unique integer), DSI-SEM(semantic structured number)
    - DSI-QG



Offline Experimental results

- **Main results**

Table 3: Experimental results on the NQ 100K dataset. ∗, † and ‡ indicate statistically significant improvements over the best performing generative retrieval baseline DSI-QG, BM25, and RepBERT, respectively ($p \leq 0.05$).

Methods	MRR@3	MRR@20	Hits@1	Hits@10
BM25	0.1846	0.1873	0.1742	0.2111
RepBERT	0.3254	0.3339	0.2993	**0.5042**
DSI-ARB	0.2224	0.2684	0.2617	0.3246
DSI-SEM	0.2516	0.2801	0.2699	0.3427
DSI-QG	0.3131	0.3220	0.2903	0.3869
SE-DSI$_{Doc}$	0.2916†	0.3001†	0.2700†	0.3627†
SE-DSI$_{Random}$	0.3046†	0.3160†	0.2866†	0.3709†
SE-DSI$_{Lead}$	0.3224†	0.3329†	0.3078*†	0.4087*†
SE-DSI$_{Sum}$	**0.3511***†‡	**0.3644***†‡	**0.3383***†‡	0.4555*†

Table 2: Experimental results on the MS MARCO dataset. ∗, † and ‡ indicate statistically significant improvements over the best performing generative retrieval baseline DSI-QG, BM25, and RepBERT, respectively ($p \leq 0.05$).

Methods	MS MARCO 10K				MS MARCO 100K				MS MARCO Full			
	MRR@3	MRR@20	Hits@1	Hits@10	MRR@3	MRR@20	Hits@1	Hits@10	MRR@3	MRR@20	Hits@1	Hits@10
BM25	0.4049	0.4230	0.3760	0.5866	0.3815	0.3700	0.4846	0.5363	0.1784	0.2168	0.1186	0.4358
RepBERT	0.4304	**0.4776**	0.4070	0.5874	0.4191	0.4459	0.4917	0.6195	0.2671	0.3078	0.1930	**0.5584**
DSI-ARB	0.1069	0.1274	0.1087	0.1377	0.1153	0.1176	0.1187	0.1180	0.1053	0.1079	0.1022	0.1138
DSI-SEM	0.2096	0.2152	0.2045	0.2392	0.2103	0.2196	0.2054	0.2544	0.1331	0.1479	0.1092	0.1678
DSI-QG	0.4237	0.4497	0.3831	0.5913	0.3997	0.4233	0.3515	0.5703	0.2277	0.2312	0.1980	0.2805
SE-DSI$_{Doc}$	0.2559	0.2631	0.2360	0.3205	0.4686*†‡	0.4757*†‡	0.4360*	0.5427	0.2429*†	0.2516*†	0.2036†	0.3347*
SE-DSI$_{Random}$	0.4217†	0.4425†	0.3725	0.5837	0.4693*†‡	0.4819*†‡	0.4320*	0.5774†	0.2577*†	0.2616*†	0.2161*†‡	0.3561*
SE-DSI$_{Lead}$	0.4343*†	0.4582†	0.3876†	**0.6063***†‡	0.5171*†‡	0.5314*†‡	0.4680*	0.6478*†‡	0.2779*†‡	0.2845*†	0.2381*†‡	0.3597*
SE-DSI$_{Sum}$	**0.4377***†	0.4567†	**0.4074***†	0.5830	**0.5900***†‡	**0.6092***†‡	**0.5347***†‡	**0.7528***†‡	**0.3022***†‡	**0.3463***†‡	**0.2609***†‡	0.4002*

SE-DSI$_{Lead}$ and SE-DSI$_{Sum}$ can perform significantly better than strong baseline solutions

Offline Experimental results

- **Analysis on rehearsal contents**

Table 6: Impact of different RCs on MS MARCO 100K. * indicates statistically significant improvements over the best performing variant w/ Doc+Psg ($p \leq 0.05$).

Methods	MRR@3	MRR@20	Hits@1	Hits@10
w/ Document	0.4686	0.4757	0.4360	0.5427
w/ Sentence	0.4326	0.4520	0.3813	0.5930
w/ Passage	0.3061	0.3143	0.2799	0.3781
w/ Doc+Sent	0.4702	0.4611	0.4844	0.6140
w/ Doc+Psg	0.4895	0.500	0.4503	0.5884
SE-DSI$_{Sum}$	**0.5900***	**0.6092***	**0.5347***	**0.7528***

SE-DSI$_{Sum}$ achieves the best results, again indicating that our method learning with the underlined important contents of the documents can comprehensively encode the documents, and further contribute to the retrieval.

Offline Experimental results

- **Zero-shot setting**
  - **only performing indexing without the retrieval task**

Table 5: Experimental results of zero-shot retrieval settings on MS MARCO 100K and NQ 100K. ∗ indicates statistically significant improvements over the best performing baseline DSI-QG ($p \leq 0.05$).

Methods	MS MARCO 100K				NQ 100K			
	MRR@3	MRR@20	Hits@1	Hits@10	MRR@3	MRR@20	Hits@1	Hits@10
DSI-ARB	0.1044	0.1135	0.1016	0.1154	0.1345	0.1380	0.1282	0.1613
DSI-SEM	0.1396	0.1545	0.1410	0.1621	0.1458	0.1507	0.1365	0.1833
DSI-QG	0.2668	0.2725	0.2468	0.3193	0.2391	0.2446	0.2019	0.2836
SE-DSI$_{Doc}$	0.2631	0.2700	0.2420	0.3258	0.1923	0.2043	0.2116	0.2660
SE-DSI$_{Random}$	0.2826*	0.2903*	0.2599*	0.3505*	0.2285	0.2320	0.2217*	0.2813
SE-DSI$_{Lead}$	0.3022*	0.3118*	0.2759*	0.3804*	0.2430	0.2517	0.2285*	0.3077*
SE-DSI$_{Sum}$	**0.4472***	**0.4326***	**0.4896***	**0.5564***	**0.2900***	**0.2947***	**0.2672***	**0.3405***

ED and RCs help the model to encode all the information about the corpus into the model parameter and SE-DSI works like a human with a knowledgeable brain.

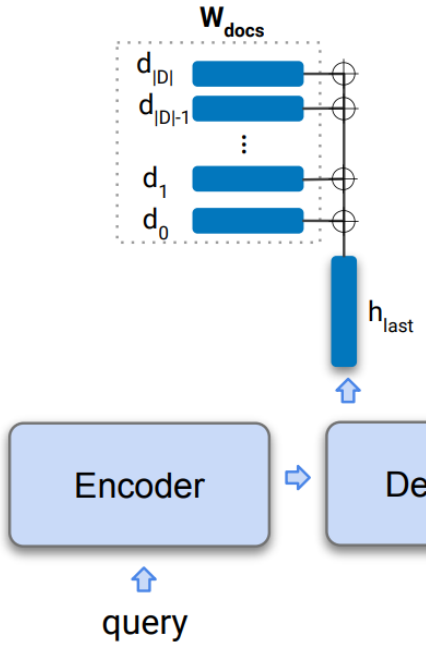
Online A/B Experiments

- **Site Retrieval task** (Applied in Baidu search)
  - The user may specify his/her information needs through a query for official sites.
  - Official sites are defined as Web pages that have been operated by universities, departments, or other administrative units

Online A/B Experiments

- **Datasets**
  - Site URL www.apple.com
  - Site name Apple official site
  - Site Domain apple.com
  - ICP record (a registration name)
  - Web page
- **Evaluation Metrics**
  - Site-level Recall@k: the predicted site URL is completely consistent with the ground-truth site URL.
  - Domain-level Recall@k: the predicted site URL and the ground-truth site URL are in the same site domain.

Online A/B Experiment



$$O = \text{Softmax}([W_{\text{docs}}]^T h_{\text{last}})$$

- **Baselines**
  - **DualEnc**
    - An Ernie-based dual-towe
    - It needs to learn a query e with (query, site attributes) pairs, where the site attributes u rd, and web page contents.
  - **SingleTow**
    - A single-tower method, including an Ernie-based encoder and a feed-forward layer, in which the weight is initialized with the site representations learned from DualEnc.
    - During training, it takes the query as input, and the output logits of the feed-forward layer are passed through a softmax function, generating a probability distribution of sites.
    - The probability of each site serves as the relevance score.

Online A/B Experiments Results

Table 8: Online A/B experimental results under the automatic evaluation. All the values are statistically significant (t-test with $p < 0.05$).

Methods	Site Level		Domain Level	
	Δ Recall@3	ΔRecall@20	Δ Recall@3	Δ Recall@20
Compared with DualEnc				
SE-DSI$_{Doc}$	+32.92%	+38.27%	+38.53%	+39.48%
SE-DSI$_{Lead}$	+36.21%	+40.93%	+41.59%	+42.11%
SE-DSI$_{Sum}$	+36.95%	+42.40%	+42.45%	+42.97%
Compared with SingleTow				
SE-DSI$_{Doc}$	+3.41%	+4.60%	+2.32%	+3.45%
SE-DSI$_{Lead}$	+6.77%	+7.32%	+5.34%	+6.13%
SE-DSI$_{Sum}$	+7.41%	+8.83%	+6.20%	+6.91%

For SE-DSI, the site representation is in the form of model parameters, making the query interact with global information, which is more flexible and deeper than explicit similarity functions.

Online A/B Experiments Results

- **Side-by-side comparison**

Table 10: Human evaluation results in terms of ΔGSB. All the values are statistically significant (t-test with $p < 0.05$).

Aspect	ΔGSB
Overall satisfaction	+2.99%
High-quality and authority	+11.52%

$$\Delta GSB = \frac{\#Good - \#Bad}{\#Good + \#Same + \#Bad}$$

SE-DSI has achieved significant positive gains in terms of both aspects

Online A/B Experiments

- **Inference speed**
  - Compared to DualEnc, the running speed of SE-DSI$_{Sum}$, which is proportional to the beam size, has been significantly improved by about 2.5 times.
  - The running speed of SE-DSI$_{Sum}$ is about the same as SingleTow, which classifies sites with one softmax operation.
  - **In general, the running speed of SE-DSI_{Sum} can meet the requirements of industrial applications.**

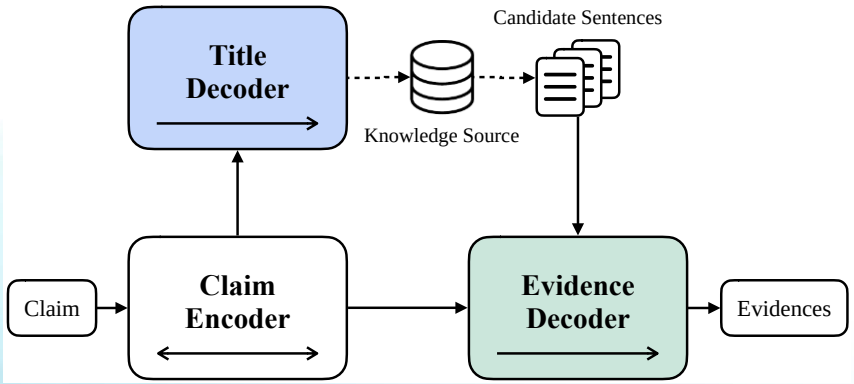
Conclusion

- Designing a proper generative model to “memorize” the whole corpus for document retrieval remains a challenge.
- Inspired by learning strategies, we have proposed SE-DSI to advance the original DSI, which takes the input of the original document augmented with RCs containing important parts and outputs the ED with explicit semantic meanings.
- The offline experimental results on several representative retrieval datasets demonstrated the effectiveness of our SE-DSI model.
- The online evaluation again verified the value of this work.

Other related work of our team

For a single task

- The first work in the field of fact-checking to use generative retrieval
- which retrieves relevant documents and evidence for claims in a generative manner, allowing for the dynamic selection of a precise set of relevant evidence for each claim.
- This not only significantly reduces memory overhead and retrieval time but also effectively enhances retrieval performance.



Model	Dev		
	P	R	F1
BM25	14.42	66.22	23.68
TF-IDF [19]	42.83	87.45	57.50
UKP-Athene [6]	35.33	**92.51**	51.13
KM [19]	44.90	83.30	58.35
NSMN [19]	52.73	88.63	66.12
DPR	55.42	89.35	68.41
RAG	62.17	91.63	74.08
GERE	**84.43**‡	78.01	**81.10**‡

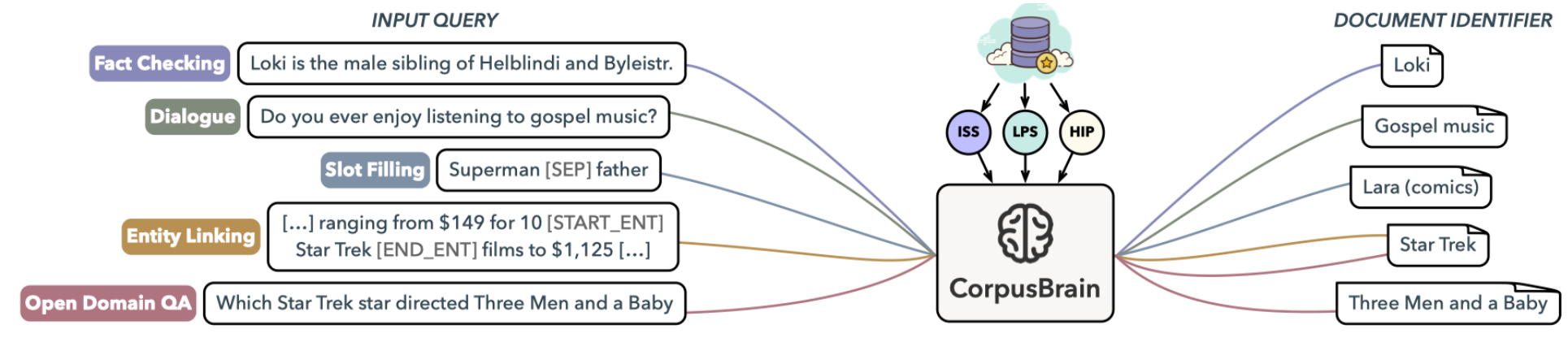
文档排序性能比较

Model	Dev			Test		
	P	R	F1	P	R	F1
TF-IDF [28]	-	-	17.20	11.28	47.87	18.26
ColumbiaNLP [2]	-	78.04	-	23.02	75.89	35.33
UKP-Athene [6]	-	87.10	-	23.61	85.19	36.97
GEAR [36]	24.08	86.72	37.69	23.51	84.66	36.80
NSMN [19]	36.49	86.79	51.38	42.27	70.91	52.96
KGAT [14]	27.29	**94.37**	42.34	25.21	**87.47**	39.14
DREAM [35]	26.67	87.64	40.90	25.63	85.57	39.45
DQN [31]	54.75	79.92	64.98	52.24	77.93	62.55
GERE	**61.43**‡	81.85	**70.18**‡	**54.30**‡	79.16	**64.41**‡

句子排序性能比较

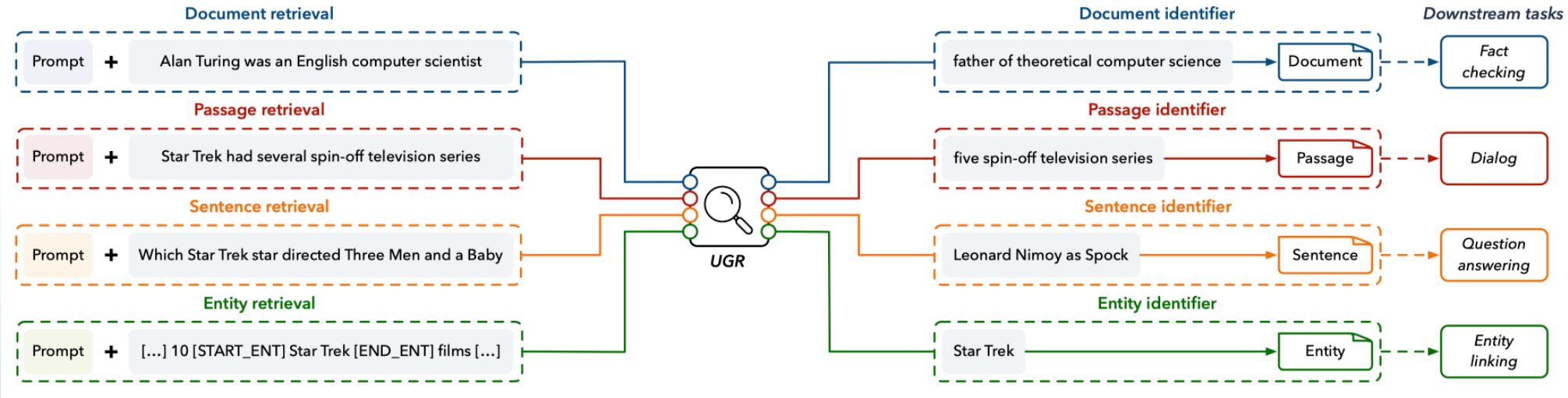
For homogeneous tasks

- CorpusBrain, the first pre-trained generative retrieval model applicable to various document retrieval tasks.
- Three self-supervised pre-training tasks (ISS, LPS, HIP) to capture query-document relevance from different perspectives.
- Outperforms existing state-of-the-art methods significantly and demonstrates strong performance in zero-resource and low-resource scenarios.



For heterogeneous tasks

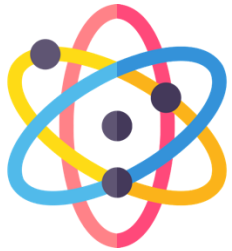
- A generative retrieval model that unifies four types of retrieval tasks: document retrieval, paragraph retrieval, sentence retrieval, and entity retrieval.
- Design a unified semantic granularity identifier based on N-Gram and leverage prompt learning to handle different retrieval tasks.
- Experimental results demonstrate the effectiveness of our approach on in-domain datasets, out-of-domain datasets, and unseen tasks.



Future work

Performance

Learning



Dynamic Corpora

Scalability

- **Encoding corpus discriminatively**
- **Fine-grained relevance modelling**

The upcoming GR tutorial

- Preliminaries: indexing, retrieval
- Docid designs
  - Pre-defined static docids： a single docid/ multiple docid
  - Learnable docids
- Training approaches
  - Static corpora: supervised learning / pre-training
  - Dynamic Corpora
- Inference strategies: constrained beam/greedy search, FM-index
- Applications: specific offline tasks, industrial applications

SIGIR-AP: November 26, 2023
The Web Conference: May 13 or 14, 2024

Q&A

Thank you!

tangyubao21b@ict.ac.cn

References

- Rethinking Search: Making Domain Experts out of Dilettantes
- AUTOREGRESSIVE ENTITY RETRIEVAL
- Gere: Generative evidence retrieval for fact verification
- Autoregressive search engines: Generating substrings as document identifiers
- Corpusbrain: Pre-train a generative retrieval model for knowledge-intensive language tasks
- Transformer memory as a differentiable search index
- Bridging the Gap Between Indexing and Retrieval for Differentiable Search Index with Query Generation