EMNLP 2023

GLEN: Generative Retrieval via Lexical Index Learning

Sunkyung Lee* Minjin Choi* Jongwuk Lee

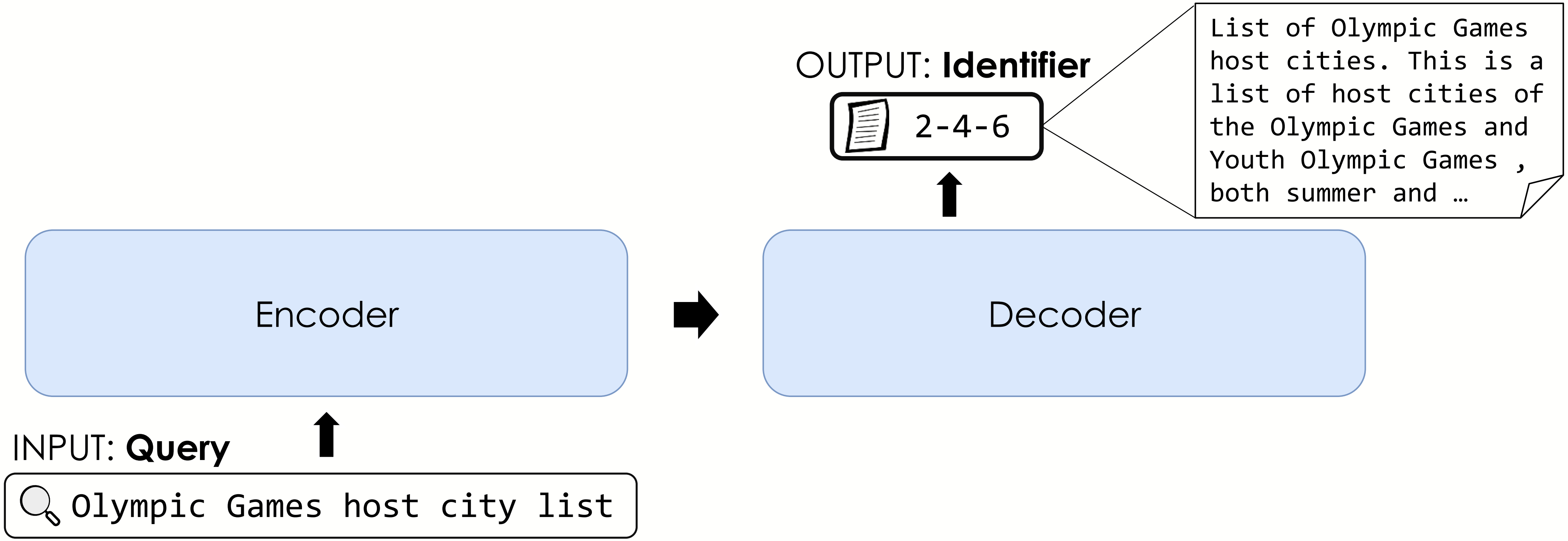
Sungkyunkwan University (SKKU), Republic of Korea

* equal contribution

Motivation

Generative Document Retrieval

- **Generative retrieval generates an identifier of a relevant document for a given query.**
  - Model parameters encode all information of the corpus, enabling end-to-end optimization.
  - The memory and computational cost of the index structure is reduced.



OUTPUT: **Identifier**

2-4-6

List of Olympic Games host cities. This is a list of host cities of the Olympic Games and Youth Olympic Games , both summer and …

Encoder

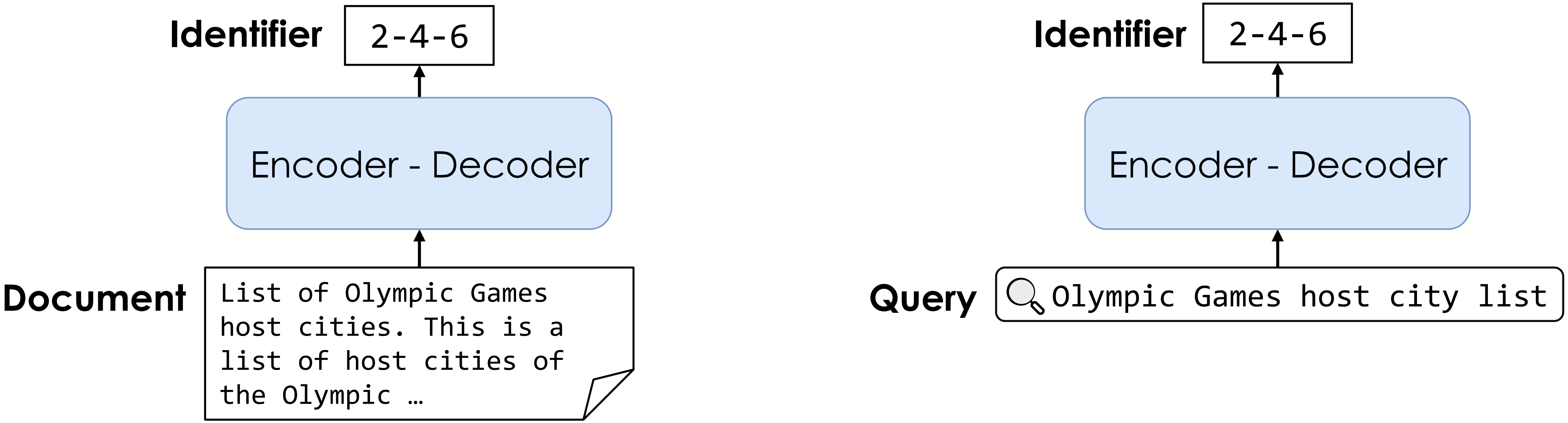
Decoder

INPUT: **Query**

Olympic Games host city list

Existing Methods

- **DSI is the first work to parameterize a retrieval system with a single transformer model.**
  - It assigns random or semantic numbers for each document and learns the assignment.



Identifier

2-4-6

Encoder - Decoder

Document

List of Olympic Games host cities. This is a list of host cities of the Olympic …

Identifier

2-4-6

Encoder - Decoder

Query

Olympic Games host city list

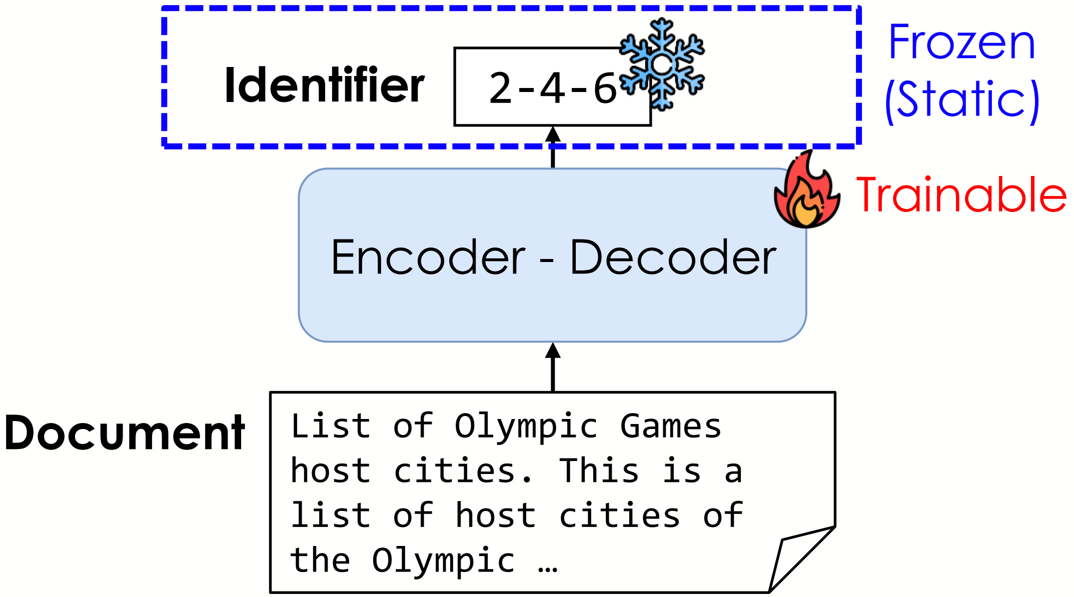
① Indexing task

② Retrieval task

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

Limitation of Existing Methods

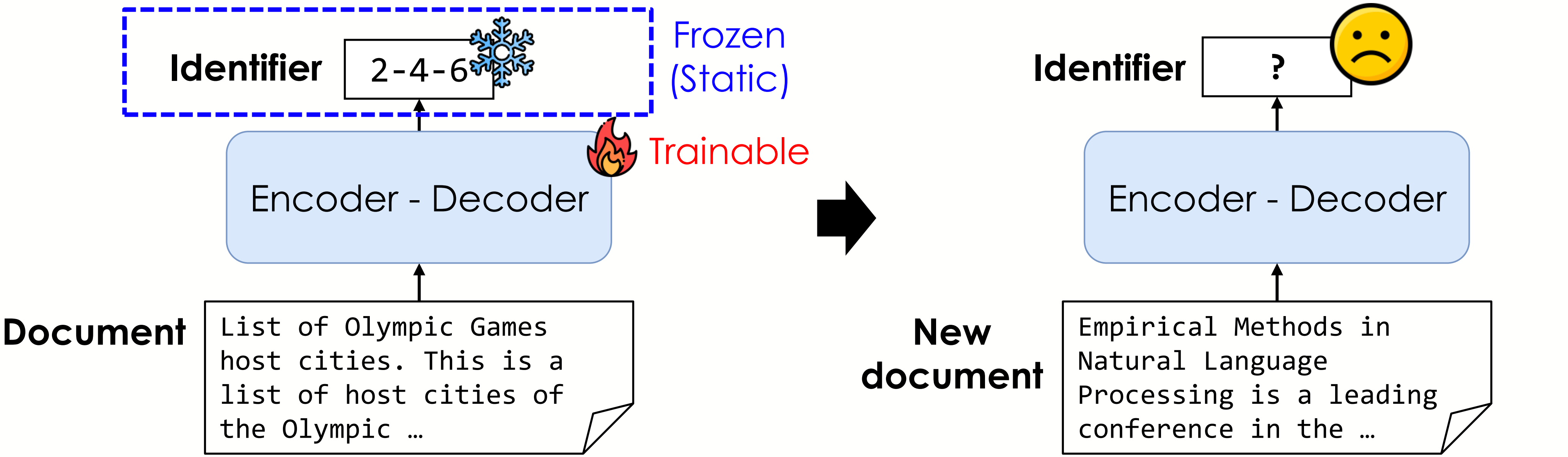
- **Most existing models pre-define static document identifiers, but they are difficult to generalize to new documents.**
  - The static identifiers can be random numbers, topic hierarchies, titles, or URLs.



```
List of Olympic Games
host cities. This is a
list of host cities of
the Olympic …
```

Limitation of Existing Methods

- **Most existing models pre-define static document identifiers, but they are difficult to generalize to new documents.**
  - The static identifiers can be random numbers, topic hierarchies, titles, or URLs.



List of Olympic Games host cities. This is a list of host cities of the Olympic …

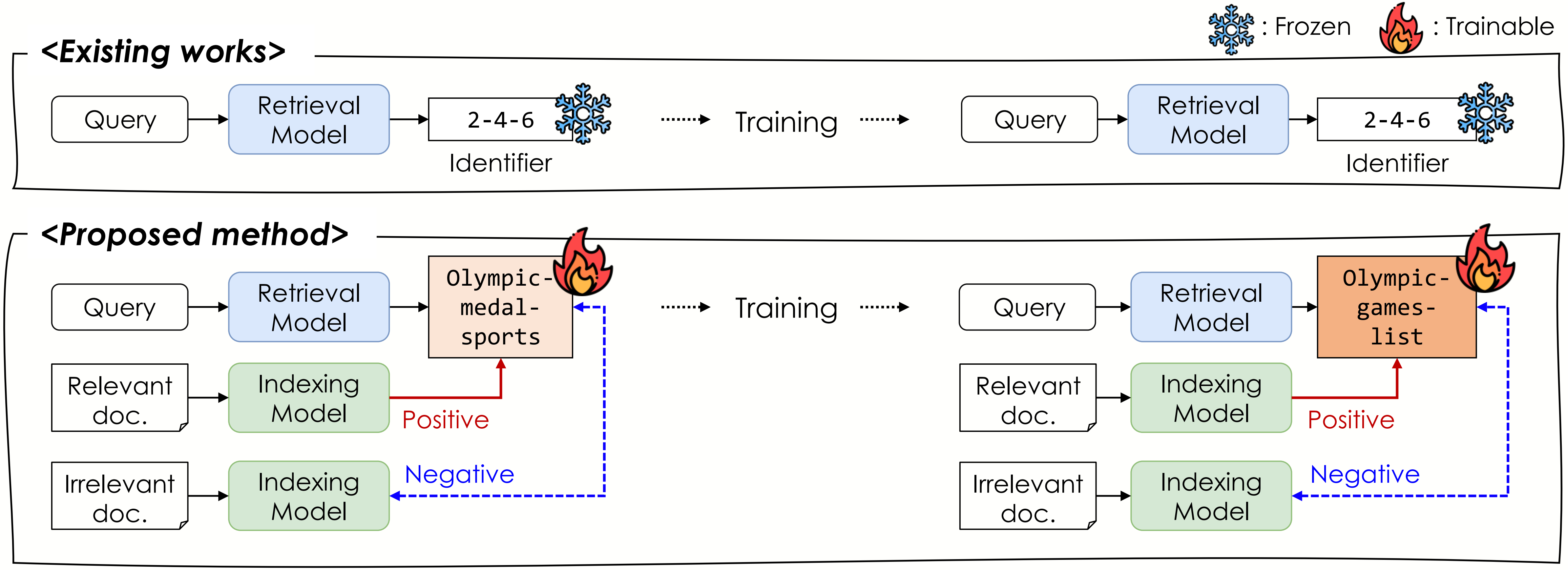
Empirical Methods in Natural Language Processing is a leading conference in the …

Research Question

How can we learn appropriate document identifiers for generative retrieval?

Our Solution: Lexical Index Learning

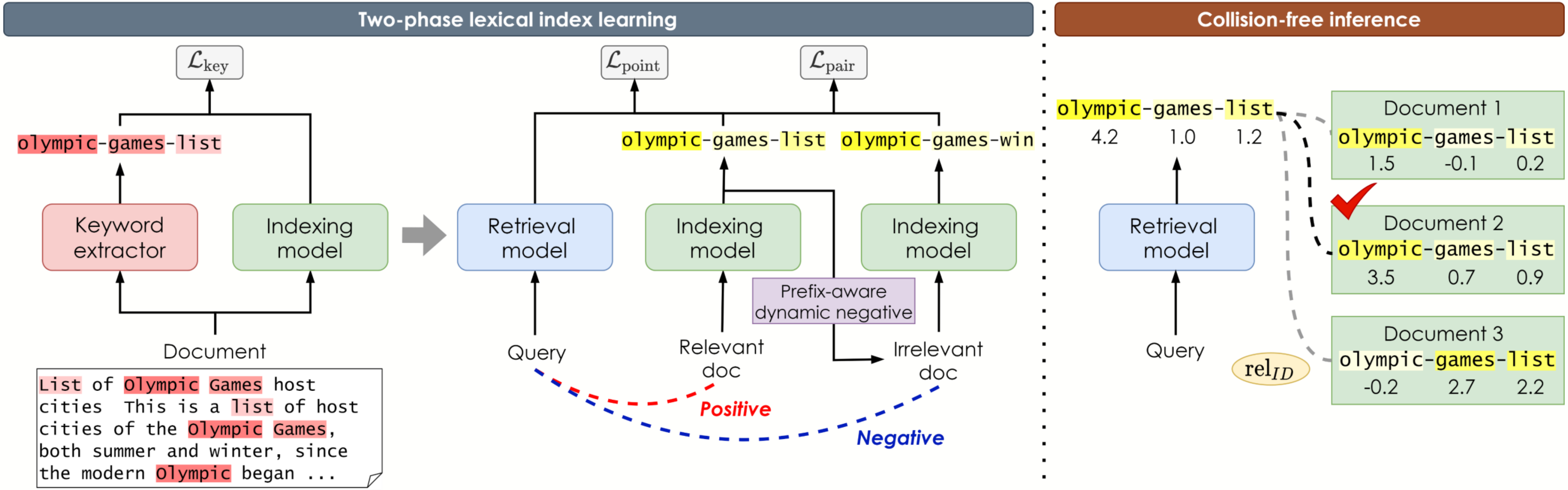
- **We devised a lexical index learning to dynamically learn identifiers considering query-document relevance.**
  - Namely GLEN (**G**enerative Retrieval via **LE**xical I**N**dex Learning)



Proposed Method

Overview of GLEN

- For training, GLEN exploits a dynamic lexical identifier using a two-phase lexical index learning to effectively learn relevance signals.
- For inference, GLEN utilizes collision-free inference to efficiently rank documents.

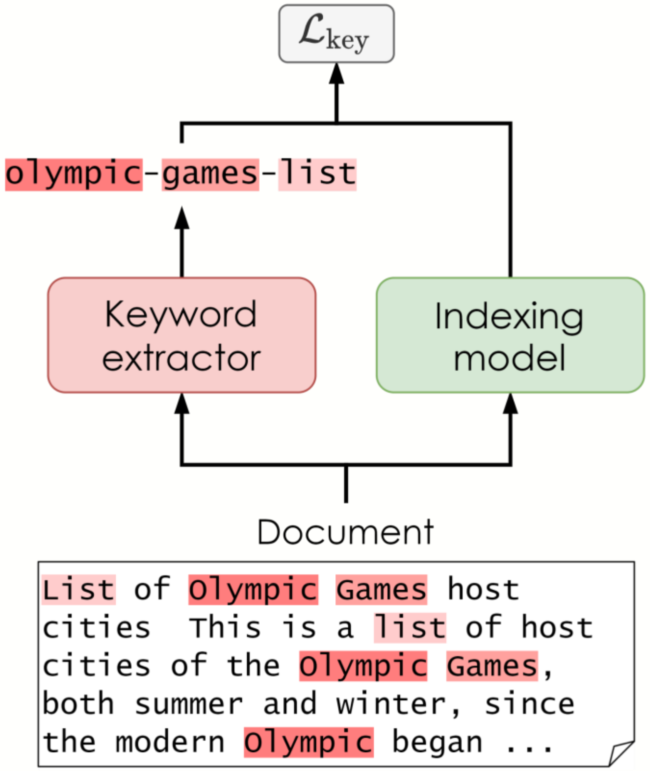


(1) Keyword-based ID assignment

(2) Ranking-based ID refinement

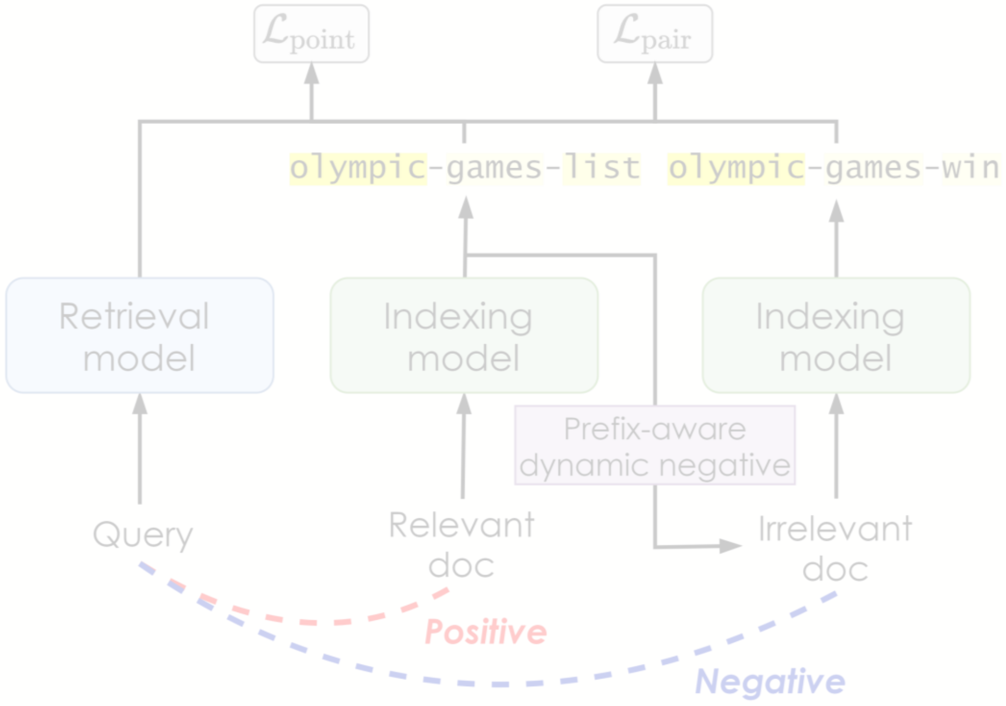
Two-phase Lexical Index Learning

- **To effectively learn the lexical index, we propose a two-phase training strategy.**
  - Phase 1: Pre-train the model using keyword-based IDs to learn the semantics of the corpus.



List of Olympic Games host cities This is a list of host cities of the Olympic Games, both summer and winter, since the modern Olympic began ...

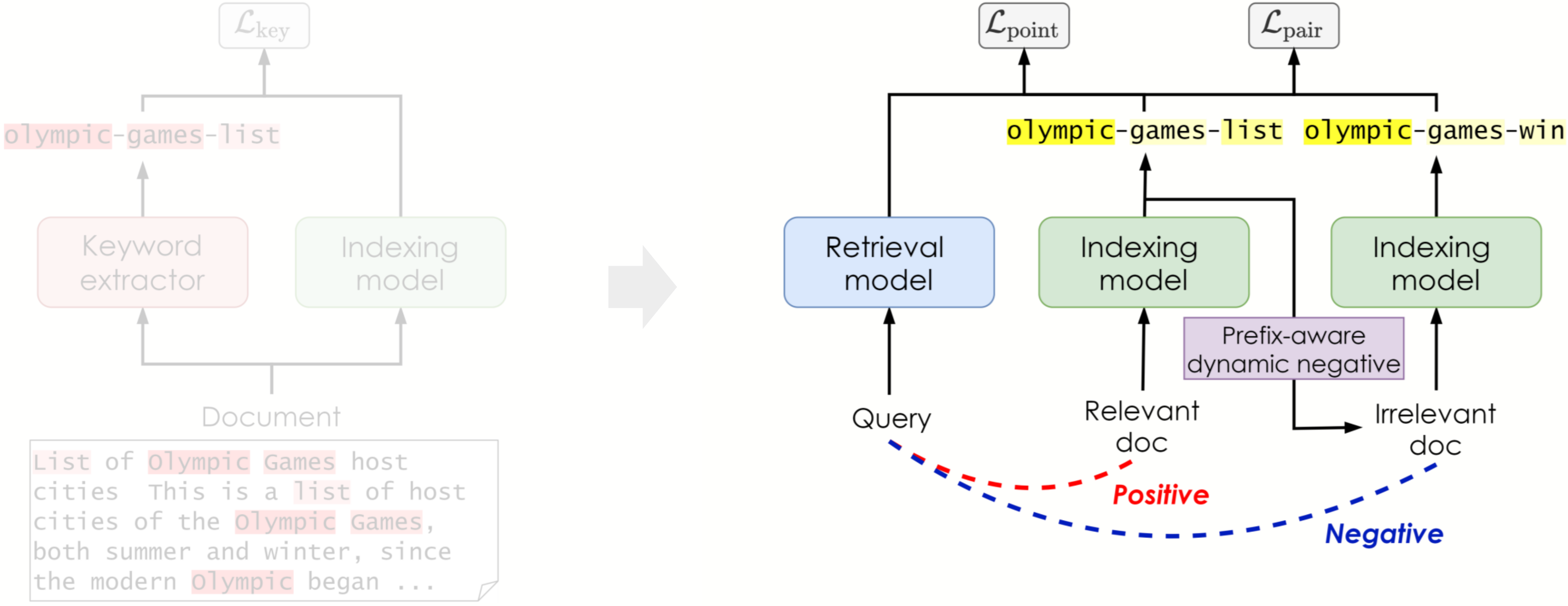
(1) Keyword-based ID assignment



(2) Ranking-based ID refinement

Two-phase Lexical Index Learning

- **To effectively learn the lexical index, we propose a two-phase training strategy.**
  - Phase 1: Pre-train the model using keyword-based IDs to learn the semantics of the corpus.
  - Phase 2: Learn ranking-based document IDs to reflect query-document relationships.


$\mathcal{L}_{\text{key}}$
olympic-games-list
Keyword extractor
Indexing model
Document
List of Olympic Games host cities This is a list of host cities of the Olympic Games, both summer and winter, since the modern Olympic began ...
$\mathcal{L}_{\text{point}}$
$\mathcal{L}_{\text{pair}}$
olympic-games-list
olympic-games-win
Retrieval model
Indexing model
Indexing model
Prefix-aware dynamic negative
Query
Relevant doc
Irrelevant doc
Positive
Negative

(1) Keyword-based ID assignment

(2) Ranking-based ID refinement

Experiments

Experimental Results: NQ320k

- GLEN achieves state-of-the-art or comparable performance compared to the best baseline on benchmark datasets.

Type	Model	Natural Questions 320K		
		R@1	R@10	MRR@100
Traditional retrieval	BM25	29.7	60.3	40.2
	DocT5Query	38.0	69.3	48.9
	DPR	50.2	77.7	59.9
	GTR-base	56.0	84.4	66.2
Generative retrieval	DSI	55.2	67.4	59.6
	DSI-QG	63.1	80.7	69.5
	NCI	66.4	85.7	73.6
	GENRET	<u>68.1</u>	88.8	75.9
	TOME	66.6	-	-
	GLEN (Ours)	69.1	<u>86.0</u>	<u>75.4</u>

*Pleases refer to the paper for more detailed results.

Experimental Results: MS MARCO & BEIR

- **GLEN yields a clear improvement over the best generative retrieval methods in large-scale corpus and zero-shot evaluation settings.**

Model	MS MARCO Dev (MRR@10)
Traditional retrieval	
BM25	18.4
DocT5Query	27.2
GTR-Base	34.8
Generative retrieval	
DSI	3.1
DSI-QG	11.8
NCI	17.4
GLEN (Ours)	20.1

Model	BEIR (nDCG@10)		
	Average	Arguana	NFCorpus
Traditional retrieval			
BM25	32.0	31.5	32.5
DocT5Query	33.9	34.9	32.8
Generative retrieval			
DSI	6.5	1.8	11.1
NCI	2.6	0.9	4.3
GENRET	12.1	12.1	12.1
GLEN (Ours)	16.8	17.6	15.9

Conclusion

Conclusion

- **We proposed a novel generative retrieval model for dynamic lexical identifiers.**
  - GLEN: **G**enerative Retrieval via **Le**xical I**n**dex Learning

- **To reflect query-document relevance, we devised a two-stage lexical index training.**

- **To resolve the identifier collision problem, we introduced collision-free inference.**

- **GLEN achieves the best or comparable performance with existing generative retrieval models on NQ320k, MS MARCO Passage Ranking, and BEIR.**

Thank you ☺
Any questions?

Email: sk1027@skku.edu

Code: https://github.com/skleee/GLEN

Paper

Code