Advances in Neural Information Retrieval (IR)

Rudra Murthy

Staff Research Scientist

rmurthyv@in.ibm.com

murthyrudra.github.io

IBM

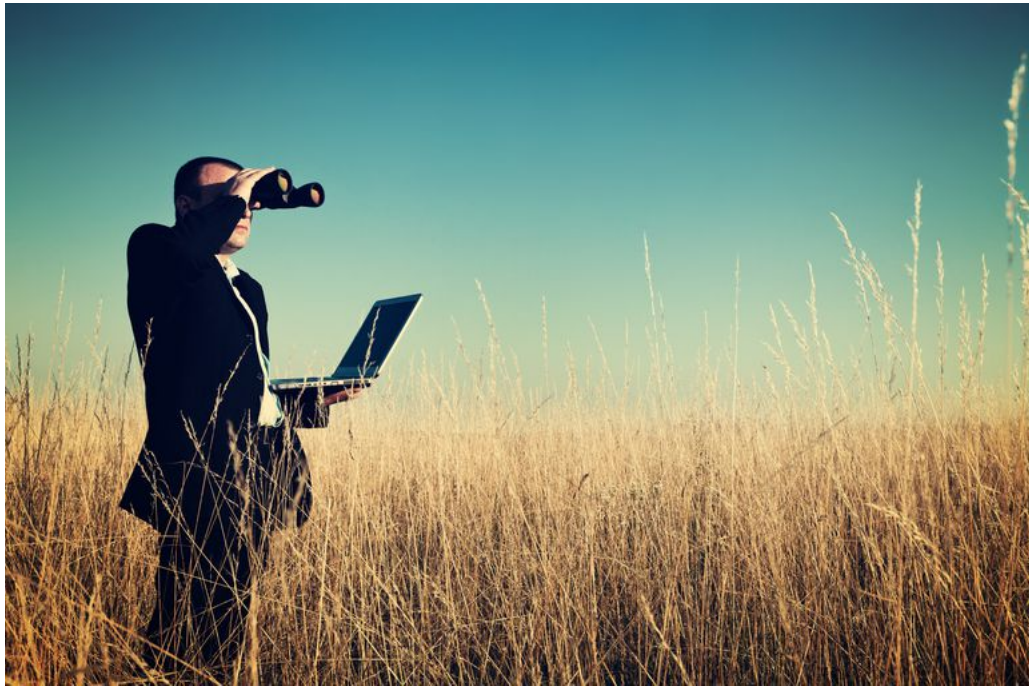
Outline

- Introduction
- Neural IR
- SDG for Retrieval

Introduction

What is Information Retrieval?

Information Retrieval is finding material of an unstructured nature that satisfies an information need from within large collections.



Reference: https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Why is IR important?

- Quickly access to relevant information from vast amounts of data
- Time efficiency: rank and filter results based on relevancy
- Enable quick decision making

IR vs Databases

Feature	Information Retrieval	Databases (SQL)
Data Nature	Unstructured (text, web pages)	Structured (tables, records)
Query	Keywords, fuzzy search	Exact matches, predefined queries
Result	Ranked documents based on relevance	Precise, structured data retrieval
Example	Google Search	Banking system retrieving customer details

Examples of IR Systems

- Web Search Engines
  - Google, Bing, DuckDuckGo, …
- E-commerce search
  - Amazon, Flipkart, …
- Library
  - IEEE, PubMed, Google Scholar, …
- Enterprise Search
  - Searching internal documents, emails, reports, …

Motivatio

Oxford University Study Shows Large Language Models (LLMs) Pose Risk to Science with False Answers

November 20, 2023 by Ali Azhar

Large Language Models (LLMs) are generative AI models that power chatbots, such as Google Bard and OpenAI's ChatGPT. There has been a meteoric rise in the use of LLMs over the last 12 months and this is indicated in several studies and surveys. However, LLMs suffer from a critical vulnerability - AI hallucination.

ChatGPT 'hallucinates.' Some researchers worry it isn't fixable.

Big Tech is pushing AI out to millions of people. But the tech still routinely makes up false answers to simple questions.

By Gerrit De Vynck

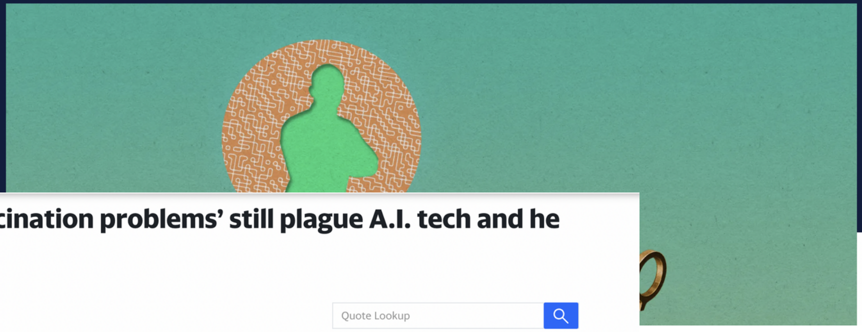
Updated May 30, 2023 at 1:27 p.m. EDT | Published May 30, 2023 at 7:00 a.m. EDT

ARTIFICIAL INTELLIGENCE

Why Big Tech's bet on AI assistants is so risky

Tech companies have not solved some of the persistent problems with AI language models.

By Melissa Heikkilä October 3, 2023



Google CEO Sundar Pichai says 'hallucination problems' still plague A.I. tech and he doesn't know why

Will Daniel

April 17, 2023 · 5 min read

Quote Lookup

TRENDING

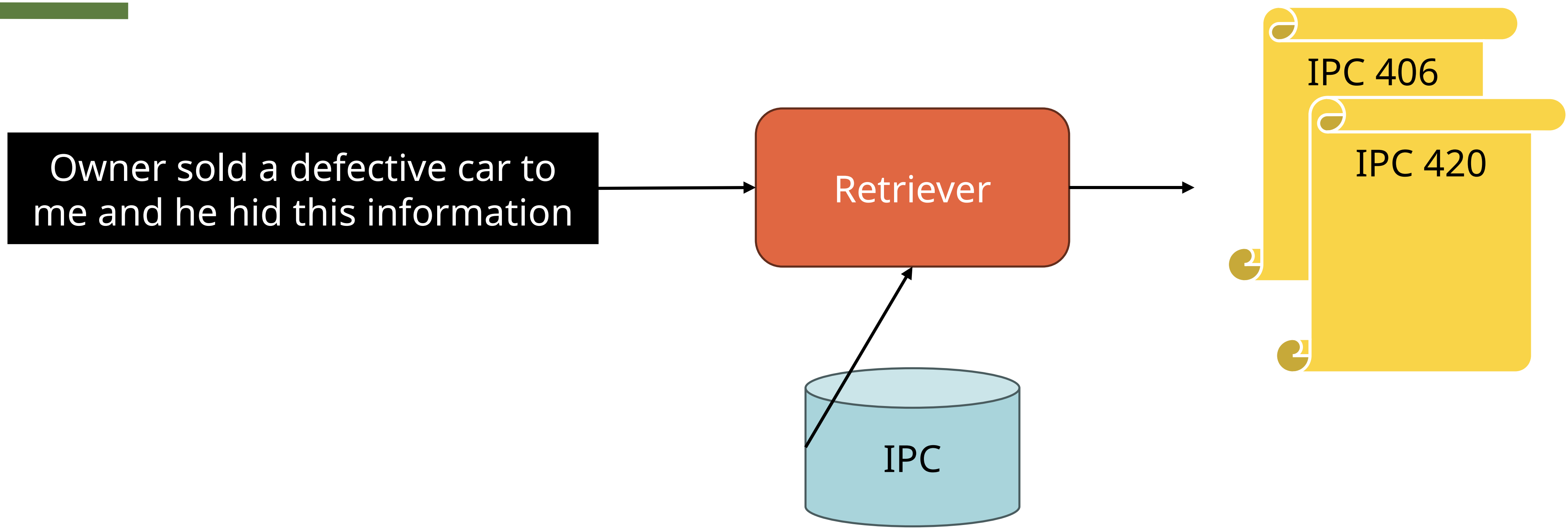
1. China's manufacturing activity slows in December in latest sign the economy is still struggling
2. Saudi sovereign wealth fund splashes cash in 2023 - report shows
3. Zelenskiy speaks of war, Putin makes passing reference in contrasting New Year speeches
4. UPDATE 2-China's Xi, US President Biden exchange congratulations on 45 years of diplomatic ties
5. INDIA RUPEE-Rupee's direction guided by Fed outlook, RBI at start of 2024

Neural IR

Source: https://web.stanford.edu/class/cs224u/slides/cs224u-neuralir-2023-handout.pdf

Code: https://github.com/murthyrudra/Information-Retrieval

Case Study


Owner sold a defective car to me and he hid this information
Retriever
IPC
IPC 406
IPC 420

VECTOR SPACE MODEL

Vector Space Model

- Treat query also as a document
- Both queries and documents are treated as vectors
- Computes similarity using cosine similarity
- This similarity metric can be used for ranking the documents

Vector Space Model

Term-Document Incidence Matrix

- Create a set of all possible words in the document collection (**Vocabulary**)
- For every term in the vocabulary, have an entry **1** if the word/term is present in the document else **0**

We can represent the query also using this term-incidence matrix

	robbery	**theft**	**…**
Document 1	1	0	..
Document 2	0	1	…
…	…	…	…

Vector Space Model

Euclidean Distance:

- If document d = (d_1, d_2, …, d_n) and q = (q_1, q_2, …, q_n)
- n is the length of vocabulary
- S(d, q) = $\sqrt{\sum_{i=0}^{n}(d_i - q_i)^2}$

Cosine Similarity:

- $\cos(\boldsymbol{d}, \boldsymbol{q}) = \frac{\boldsymbol{d} \cdot \boldsymbol{q}}{|\boldsymbol{d}||\boldsymbol{q}|} = \frac{\sum_{i=0}^{n} d_i q_i}{\sqrt{\sum_{i=0}^{n} d_i^2}\sqrt{\sum_{i=0}^{n} q_i^2}}$

- Intuitively, the larger the overlap of words between the query and document, the larger will be the similarity

Vector Space Model: Term Weighting

Consider the query

- While travelling to work today, someone hit my car and started abusing me even it wasn't my fault. What laws will help me?
- It is important to give importance to words/terms relevant for retrieval

	work	today	hit	car	abuse	fault	if	it	to	while
Query	1	1	1	1	1	1	1	1	0	1
…	…	…	…	…						

Our encoding gives importance to all words in query/document

How do we determine such TERM weights?

Vector Space Model: Term Weighting

Term Frequency

- How many times a term appears in a document
- In our example query, While travelling to work today, someone hit my car and started abusing me even if it wasn't my fault. What laws will help me?
- If the terms hit , car , abuse appears frequently, then the document is relevant for us

Vector Space Model: Term Weighting

Term Frequency

- How many times a term appears in a document
- Closed-words (Function words) are very common
  - Example: in, at , a , an , the
- How do we separate the important terms from the function words?
- In our example query, While travelling to work today, someone hit my car and started abusing me even if it wasn’t my fault. What laws will help me?
- The irrelevant terms in the above query are
  - while, if, my, …

Vector Space Model: Inverse Document Weighing

Inverse Document Frequency

- Closed-words (Function words) maybe common in a document
- But they also appear in other documents too
- However, **relevant terms** appear only in that or a small subset of related documents

$\boldsymbol{df_t}$ is the document frequency of the term t

- Number of documents that contain the term t
- Higher number means less informative

IDF, $idf_t = log_{10} \frac{N}{df_t}$

N is the total number of documents

Vector Space Model: Inverse Document Weighing

Inverse Document Frequency

- In our example query, While travelling to work today, someone hit my car and started abusing me even if it wasn’t my fault. What laws will help me?
- Terms/words like if or while or will or even appear in many documents
- Terms like hit or abuse appear in very few documents only

Vector Space Model: Tf-IDF

The weight assigned for a particular term and a document is given by

$$w_{t,d} = \log(1 + tf_{t,d}) \times log_{10} \frac{N}{df_t}$$

High

- If a particular word appears more frequently in a document
- If the same word appears only in a subset of documents

Vector Space Model: Tf-IDF

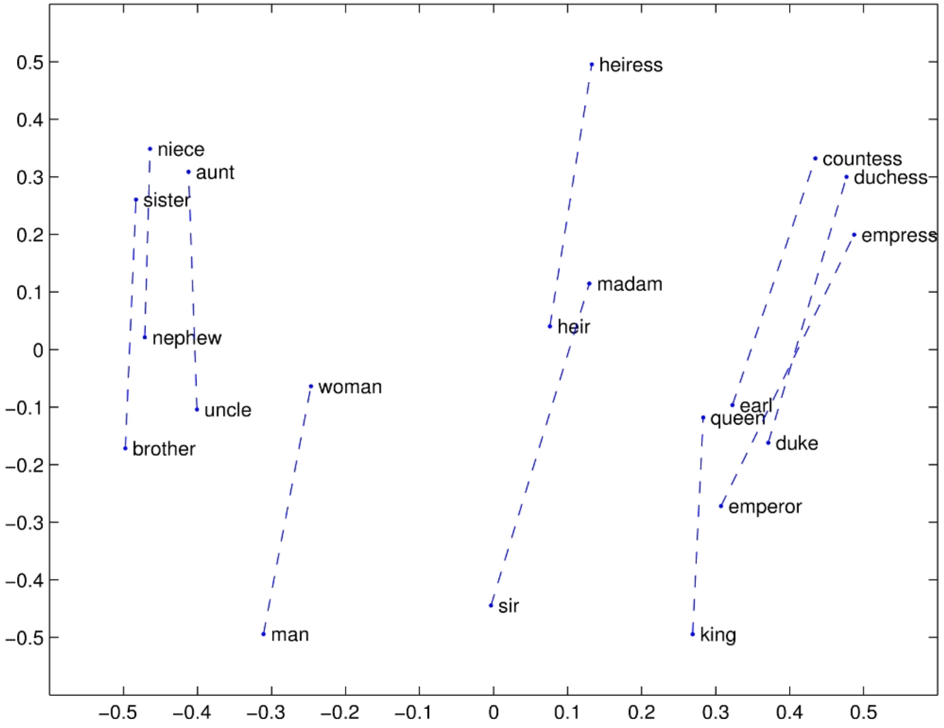
	work	today	hit	car	abuse	fault	if	it	to	while
Query	0.1	0.04	0.5	0.3	0.6	0.2	0.01	0.003	0	0.004
…	…	…	…	…						

Hypothetical query representation using tf-idf weighting

NEURAL IR

- Approaches like TF-IDF and BM25 puts too much emphasis on lexical similarity
- These retrievers might fail to capture semantic similarity

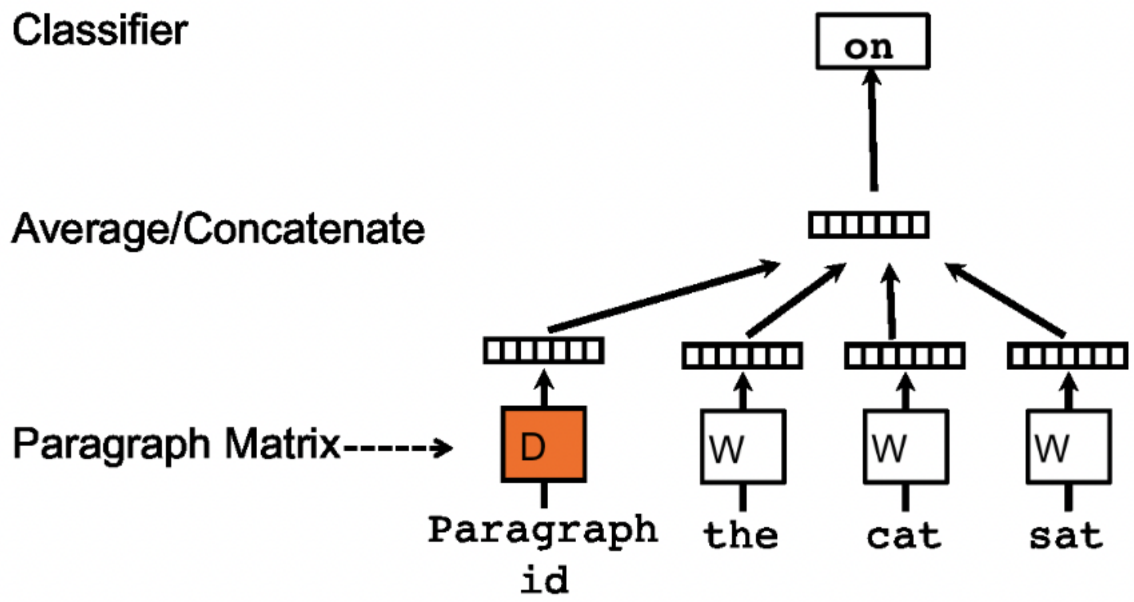
Embedding Based Retrieval



https://nlp.stanford.edu/projects/glove/

- What if we take average of word embedding of all words in the sentence/document?
- What if we use TF-IDF to determine weights, and then use these weights to take weighted average of word embeddings?

Embedding Based Retrieval



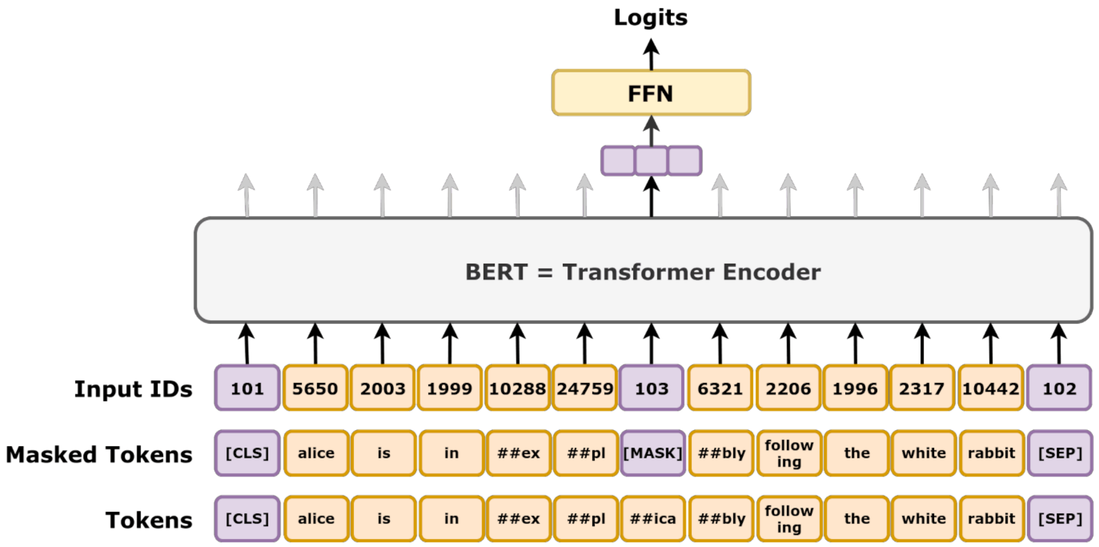
Distributed Representations of Sentences and Documents. https://arxiv.org/abs/1405.4053

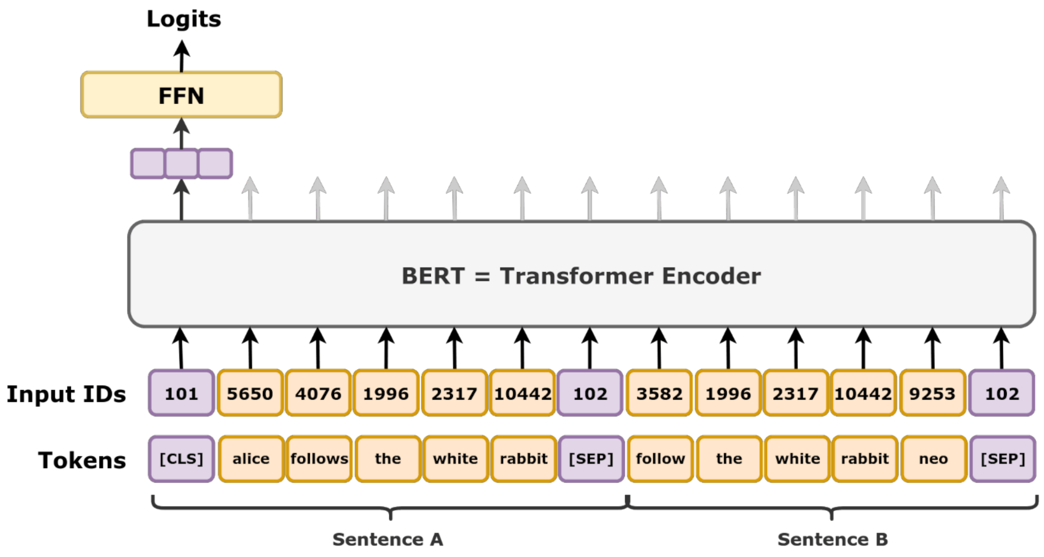
- Learn representations of sentences and paragraphs

To-Do

- Implement word embedding based query/document representation calculation
- Implement word embedding based query/document representation calculation and used TF-IDF for term weights

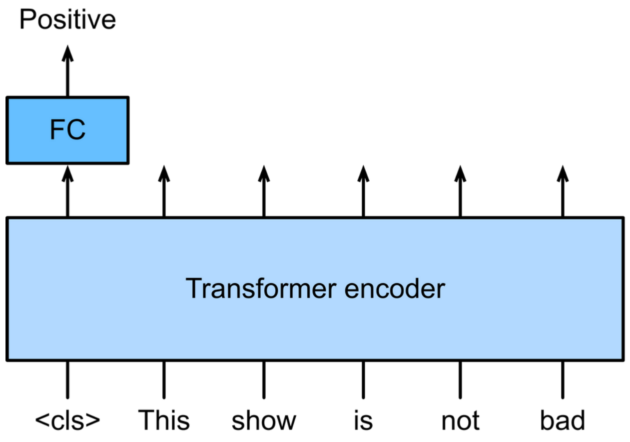
Pre-Trained language Models



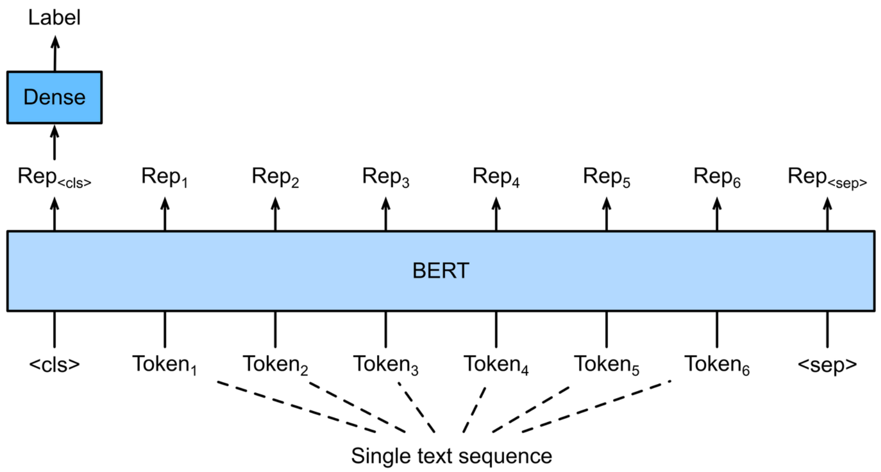


https://en.wikipedia.org/wiki/BERT_(language_model)

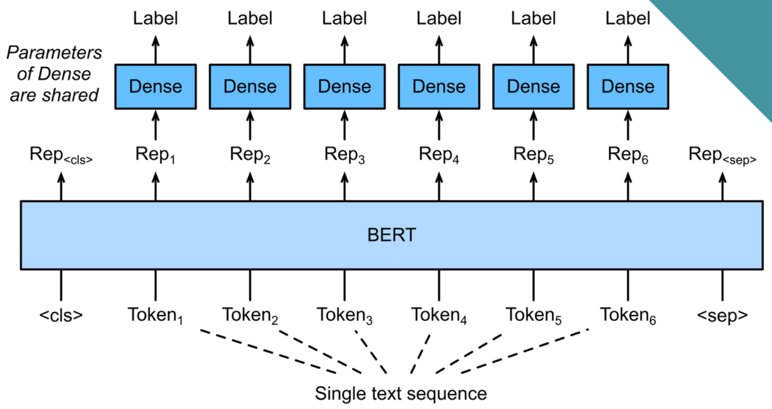
Finetuning PLM



Sentiment Analysis



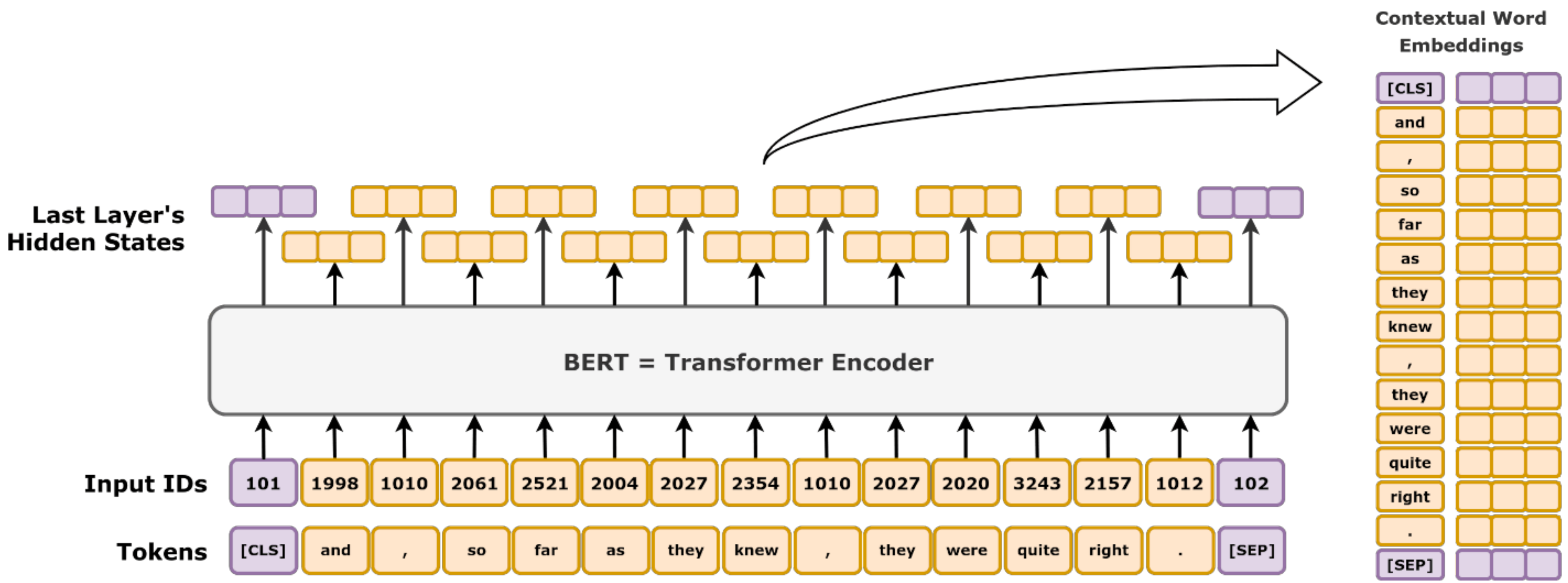
Sentence Classification



Sequence Labelling

https://en.wikipedia.org/wiki/BERT_(language_model)

PLMs



Contextual Word Embeddings
[CLS]
and
,
so
far
as
they
knew
,
they
were
quite
right
.
[SEP]
Last Layer's Hidden States
BERT = Transformer Encoder
Input IDs
101
1998
1010
2061
2521
2004
2027
2354
1010
2027
2020
3243
2157
1012
102
Tokens
[CLS]
and
,
so
far
as
they
knew
,
they
were
quite
right
.
[SEP]

Motivation

- BERT uses [CLS] as a special token in front of every sentence
- [CLS] is used for next-sentence prediction
    - What if we used [CLS] to get query/passage representation
- Alternatively, we can take the average of all token representations to form query/passage representation

To-Do

- Implement [CLS] based query/document representation from PLMs like BERT, RoBERTA, etc
- Implement average of token embeddings based query/document representation from PLMs like BERT, RoBERTA, etc

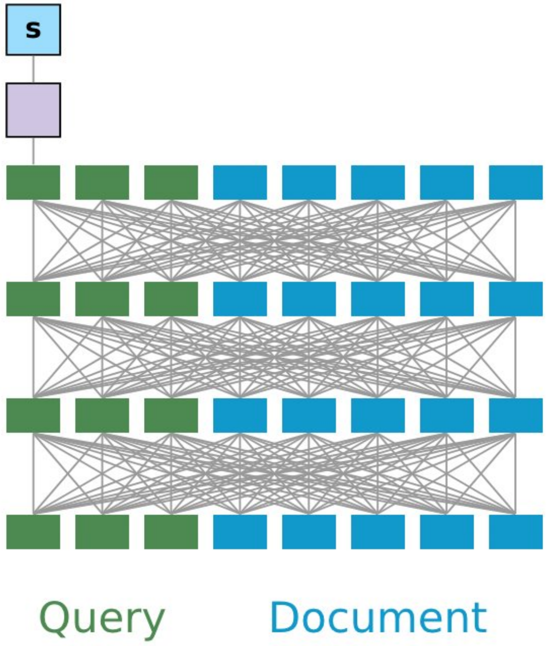
Neural Retrievers

- Cross Encoders
- Dense Passage Retrievers
- Late Interaction Models
- Sparse Retrievers

Cross Encoders

Cross-Encoders

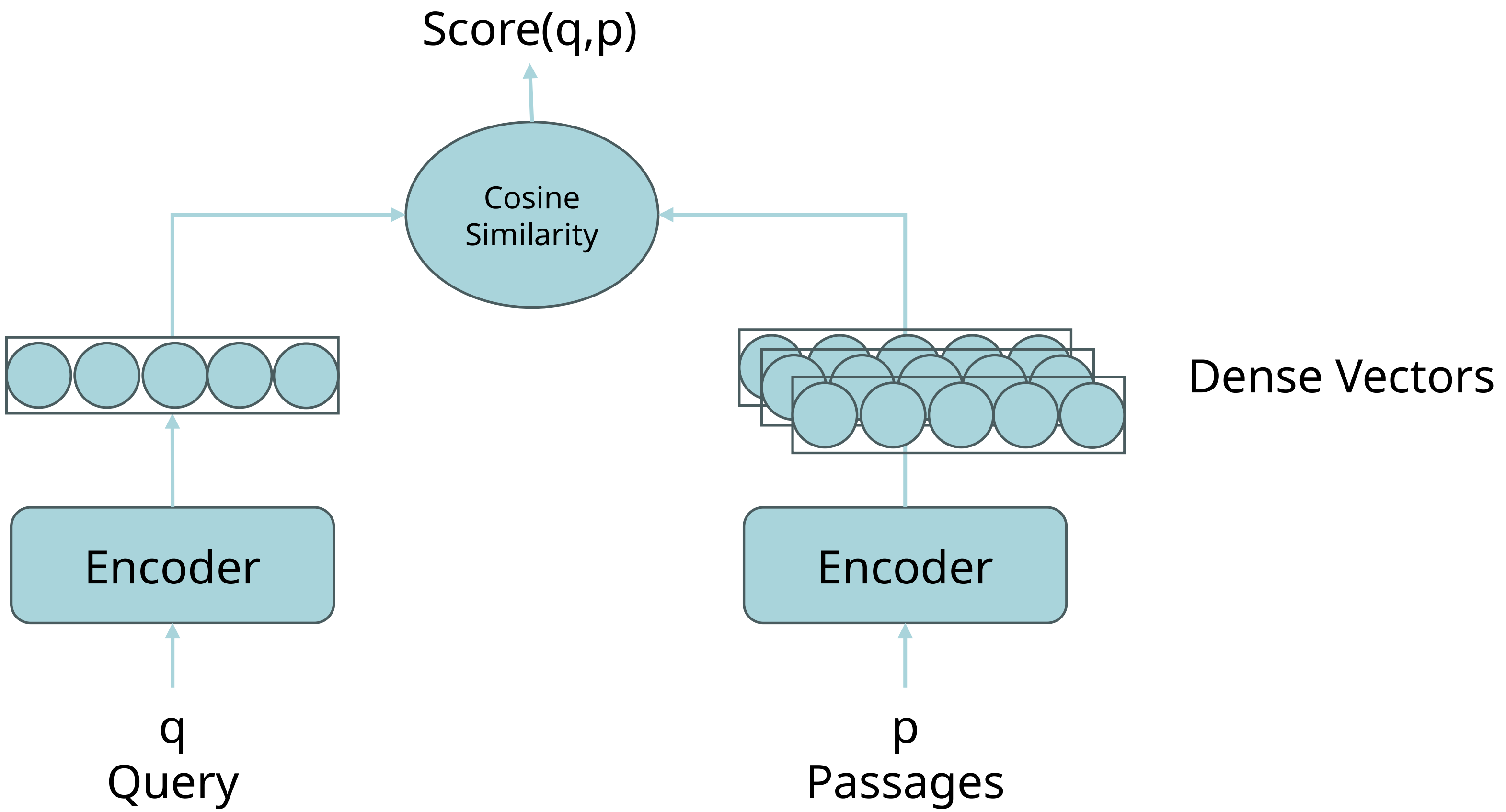
- Maximal interaction between query and document tokens
- Scalability issues



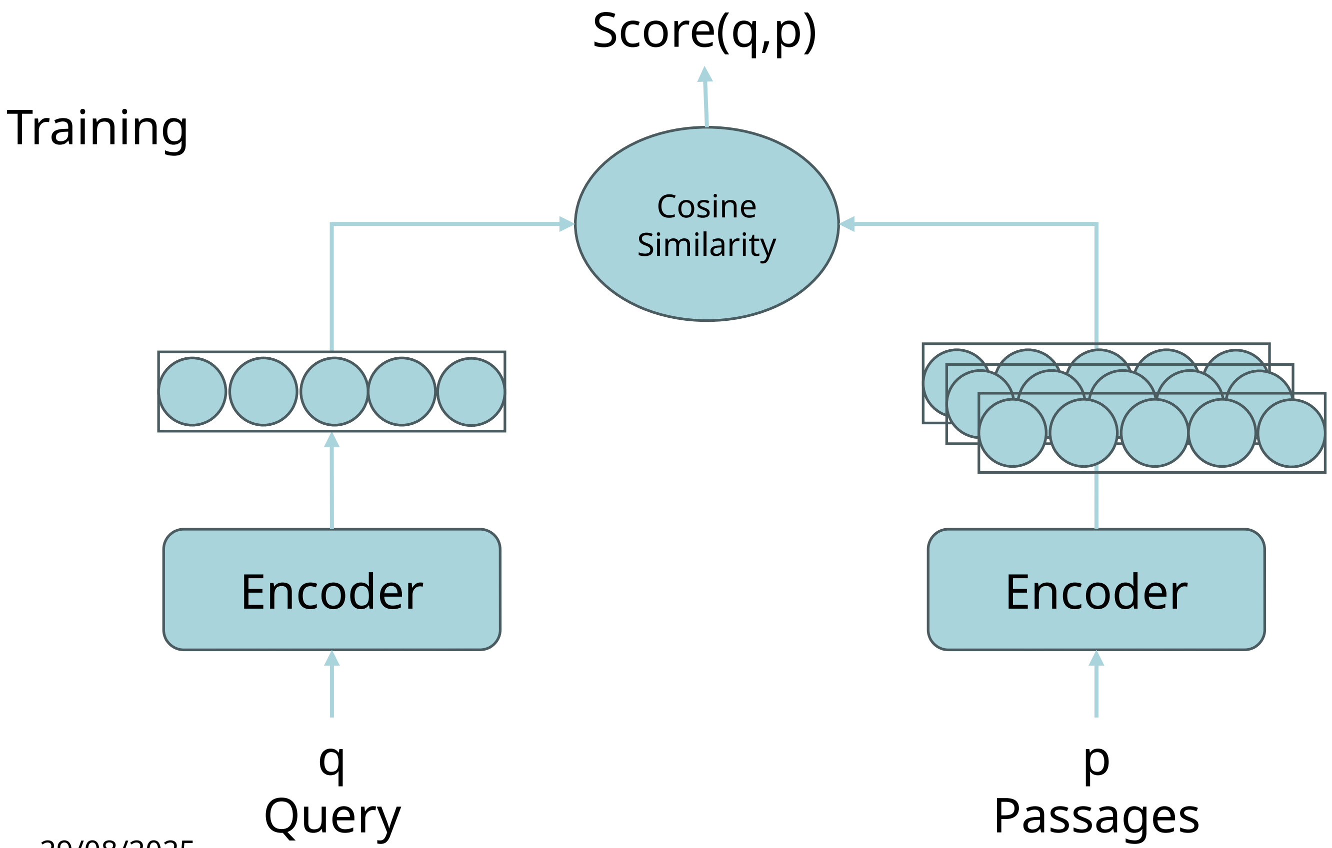
https://web.stanford.edu/class/cs224u/slides/cs224u-neuralir-2023-handout.pdf

Dense Passage Retrievers

Inferenc
e



Dense Passage Retrievers



Contrastive Learning

Given query (q), positive passage (p^+), and a set of negative passages (p^-_1, p^-_2, … p^-_n)

$L(q, p^+, p^-_1, p^-_2, \ldots p^-_n) =$

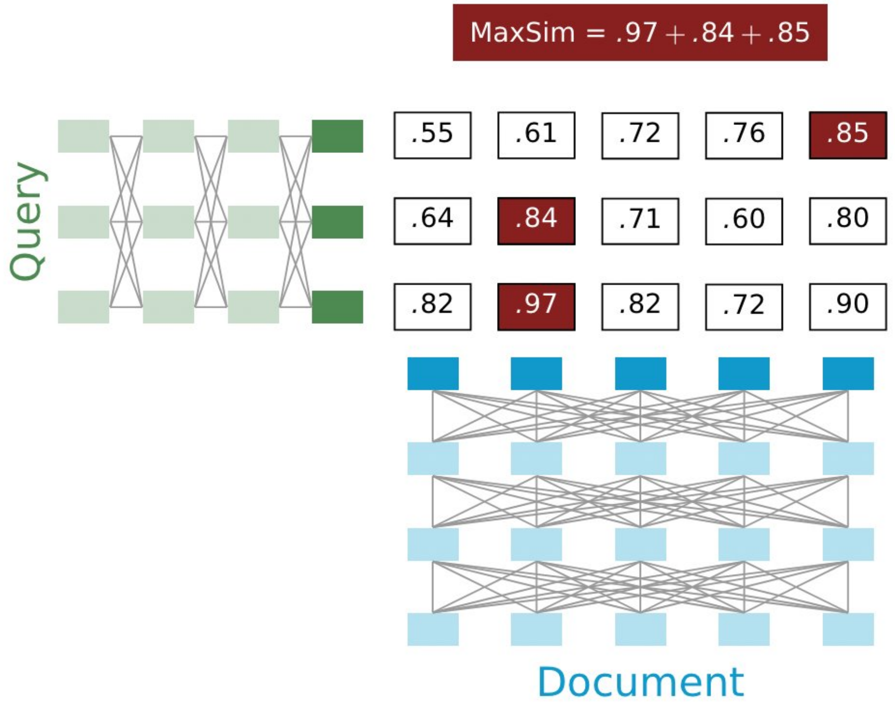
$$-\log \frac{\exp(sim(q,p^+))}{\exp(sim(q,p^+)) + \sum_{j=1}^{n} \exp(sim(q,p_j^-))}$$

Dense Passage Retrievers

- Highly Scalable
- Limited query and document interactions
- Will a d-dimensional representation be able to capture the nuances in query and documents?

ColBERT

- Late Contextual Interactions
- Every query token interacts with every passage/document token
- Have to save representation of every token in all documents



Sparse Retrievers

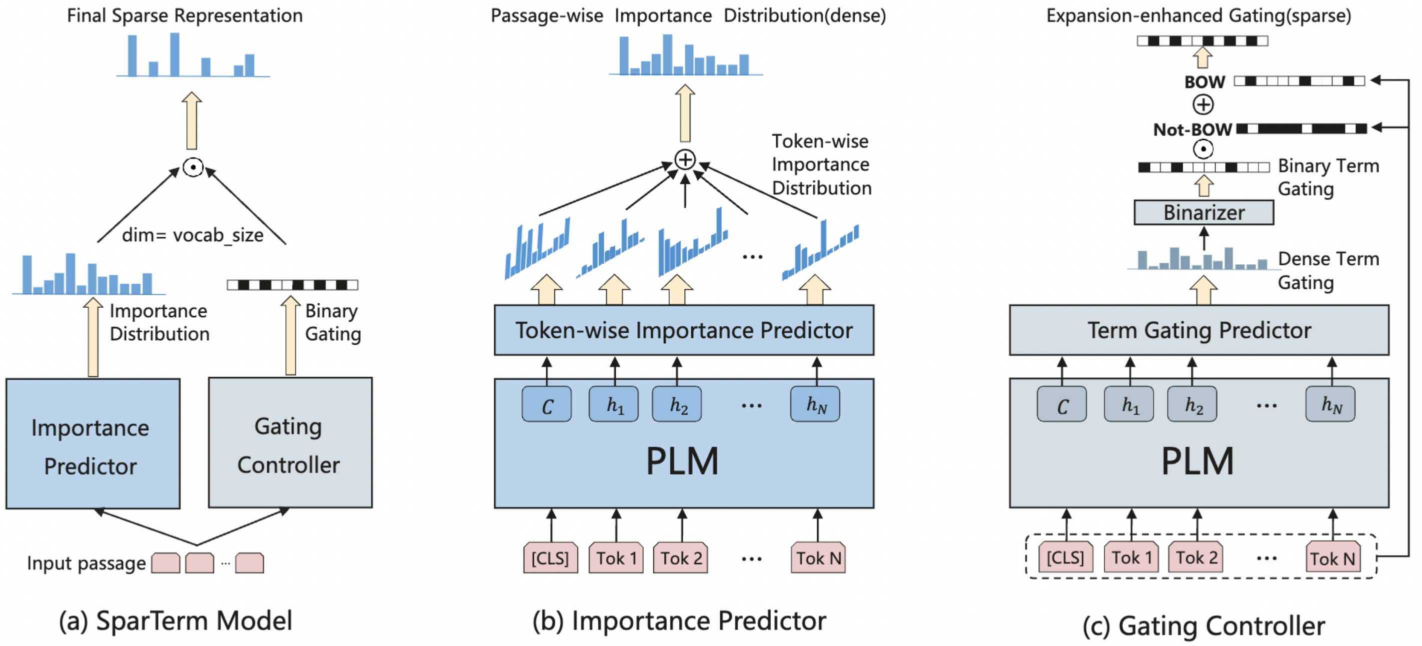


Figure 2: Model Architecture of SparTerm. Our overall architecture contains an importance predictor and a gating controller. The importance predictor generates a dense importance distribution with the dimension of vocabulary size, while the gating controller outputs a sparse and binary gating vector to control term activation for the final representation. These two modules cooperatively ensure the sparsity and flexibility of the final representation.

SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking
SparTerm: Learning Term-based Sparse Representation for Fast Text Retrieval

Training

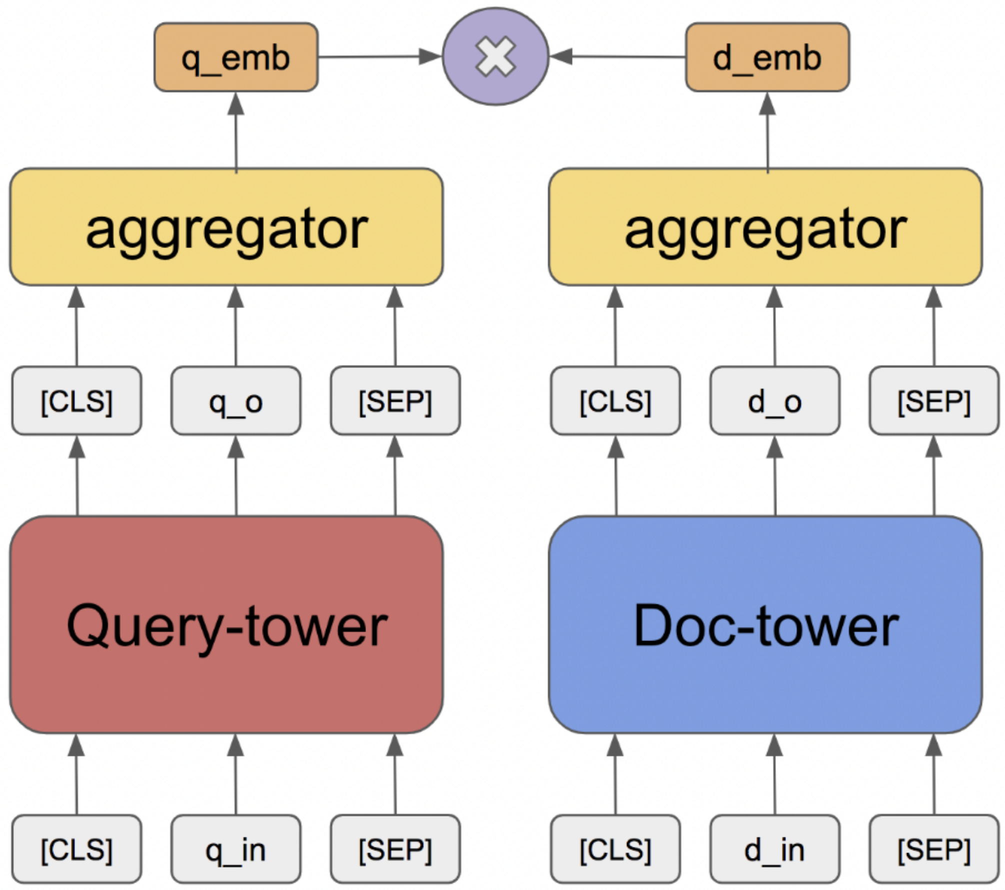
Unsupervised Pre-Training

Supervised Finetuning (Stage 1)

Supervised Finetuning (Stage 2)

Supervised Finetuning (Stage 3)

Pre-training Tasks for Embedding-based Large-scale Retrieval. ICLR 2020



Unsupervised Pre-training

Inverse Cloze Tasks

Body First Selection

Wiki Link Prediction

Supervised Fine-Tuning

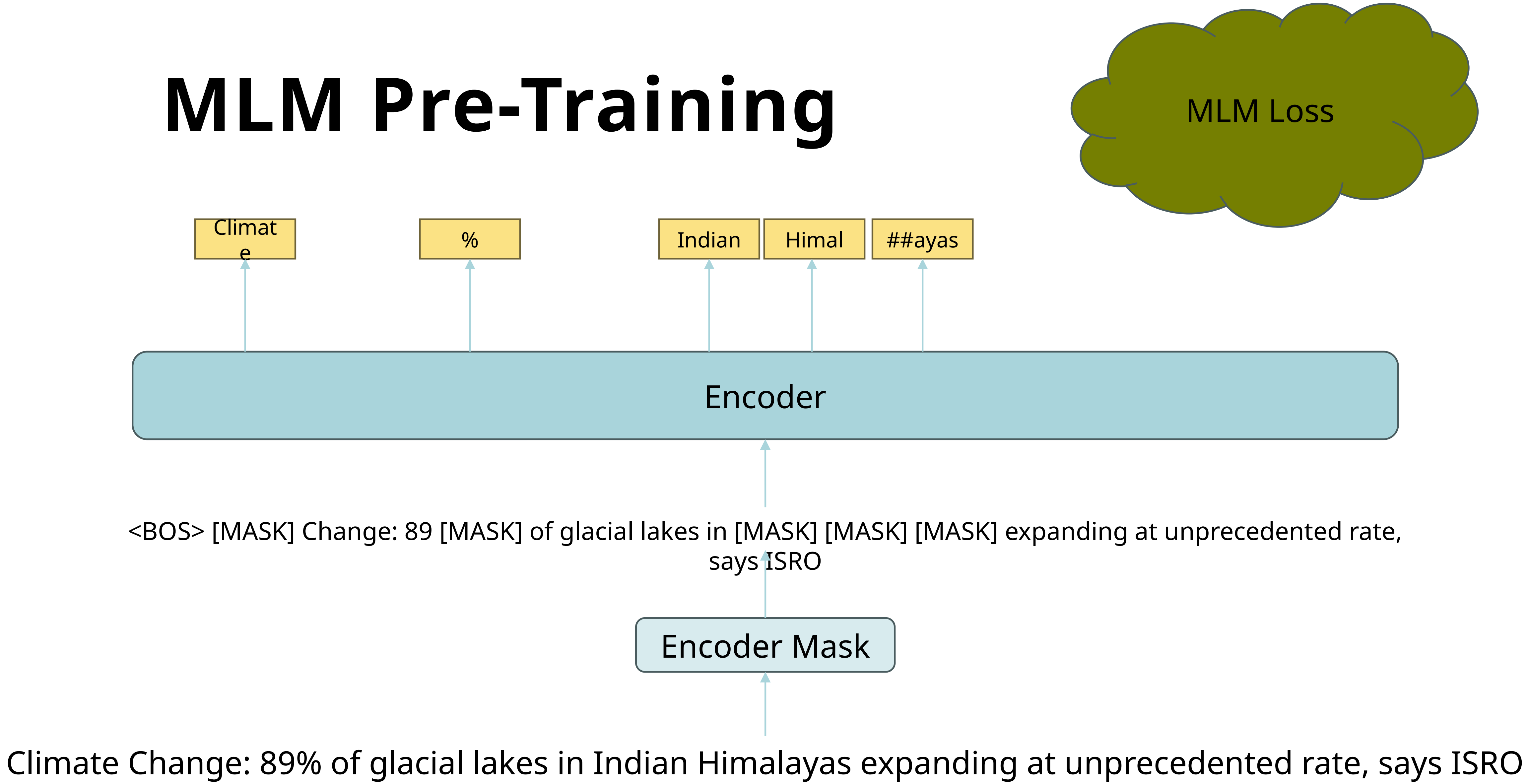
- Stage 1:
  - Use In-batch negatives with/without hard negatives mined from BM25
- Stage 2:
  - Use Stage-1 model to mine hard negatives and fine-tune a stage 2 model
- Stage 3:
  - Use knowledge distillation

IR Pretraining Strategies

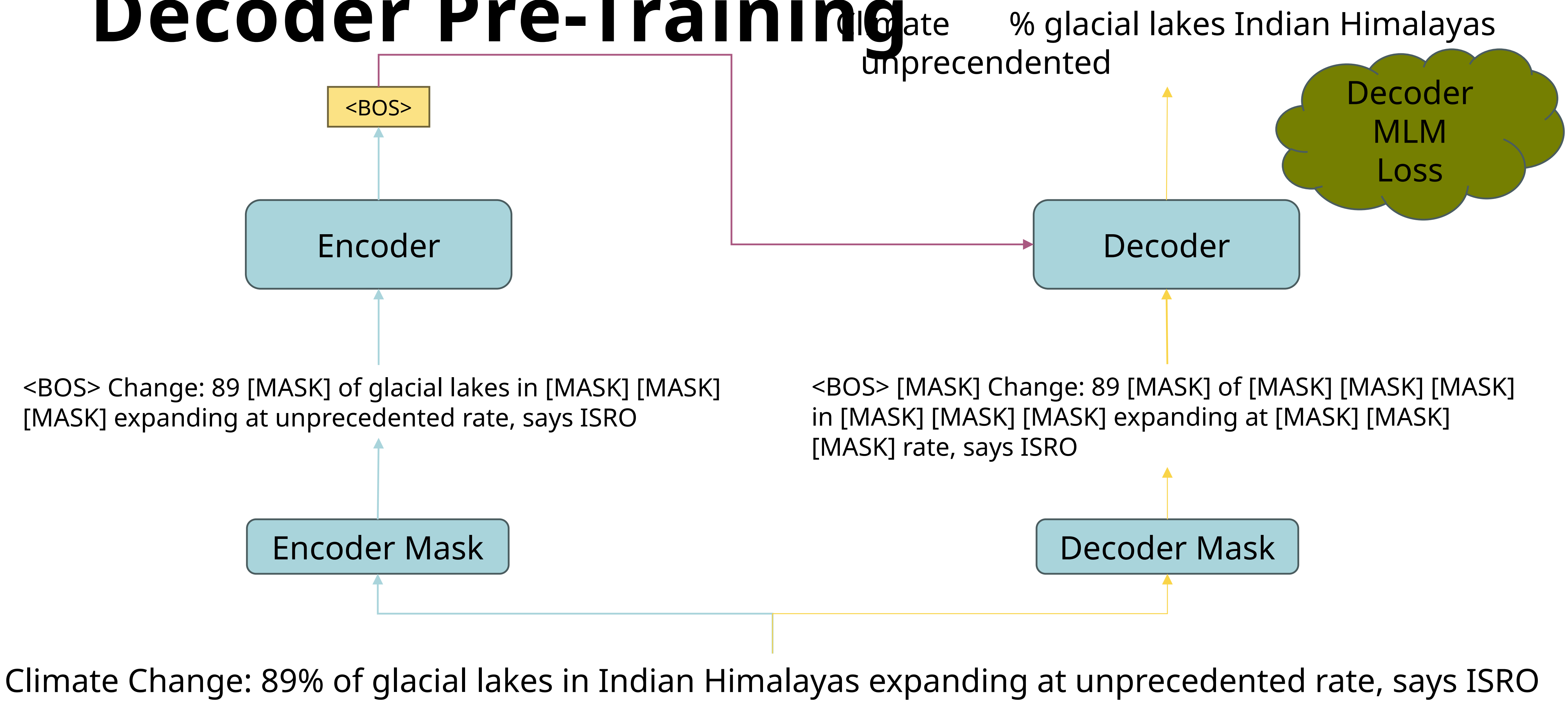
RETRO-MAE

SHITAO XIAO, ZHENG LIU, YINGXIA SHAO, AND ZHAO CAO. 2022.
RETROMAE: PRE-TRAINING RETRIEVAL-ORIENTED LANGUAGE MODELS VIA MASKED AUTO-ENCODER.
IN *PROCEEDINGS OF THE 2022 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING*, PAGES 538–548, ABU DHABI, UNITED ARAB EMIRATES. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS.

MLM Pre-Training



Decoder Pre-Training

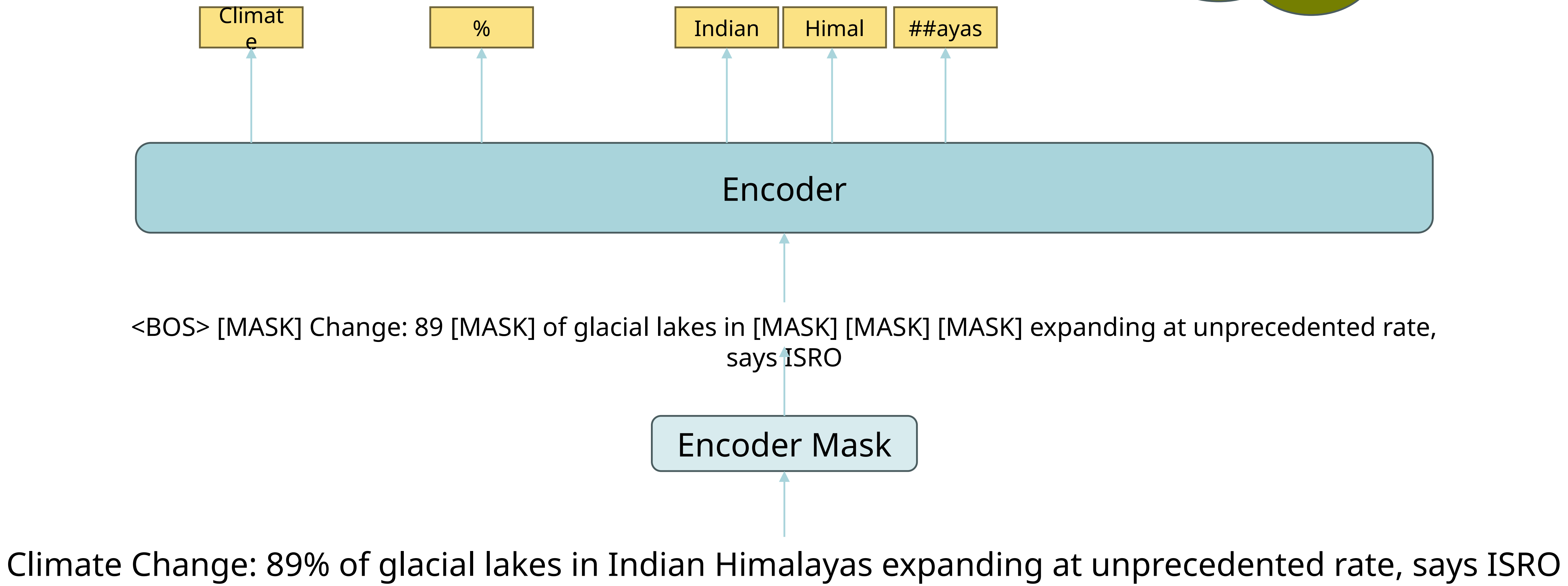


RETRO-MAE V2

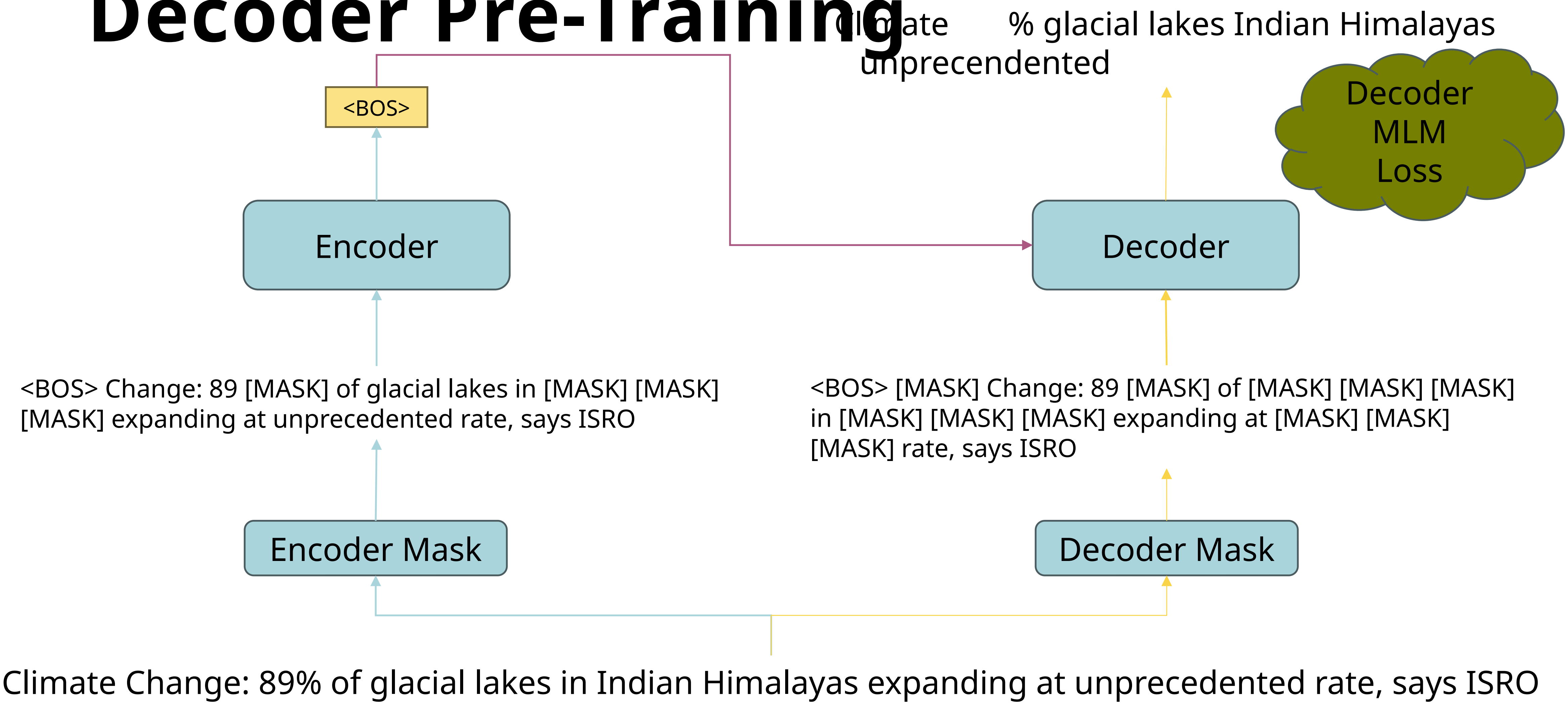
ZHENG LIU, SHITAO XIAO, YINGXIA SHAO, AND ZHAO CAO. 2023. RETROMAE-2: DUPLEX MASKED AUTO-ENCODER FOR PRE-TRAINING RETRIEVAL-ORIENTED LANGUAGE MODELS. IN *PROCEEDINGS OF THE 61ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (VOLUME 1: LONG PAPERS)*, PAGES 2635–2648, TORONTO, CANADA. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS.

PRE-TRAINING

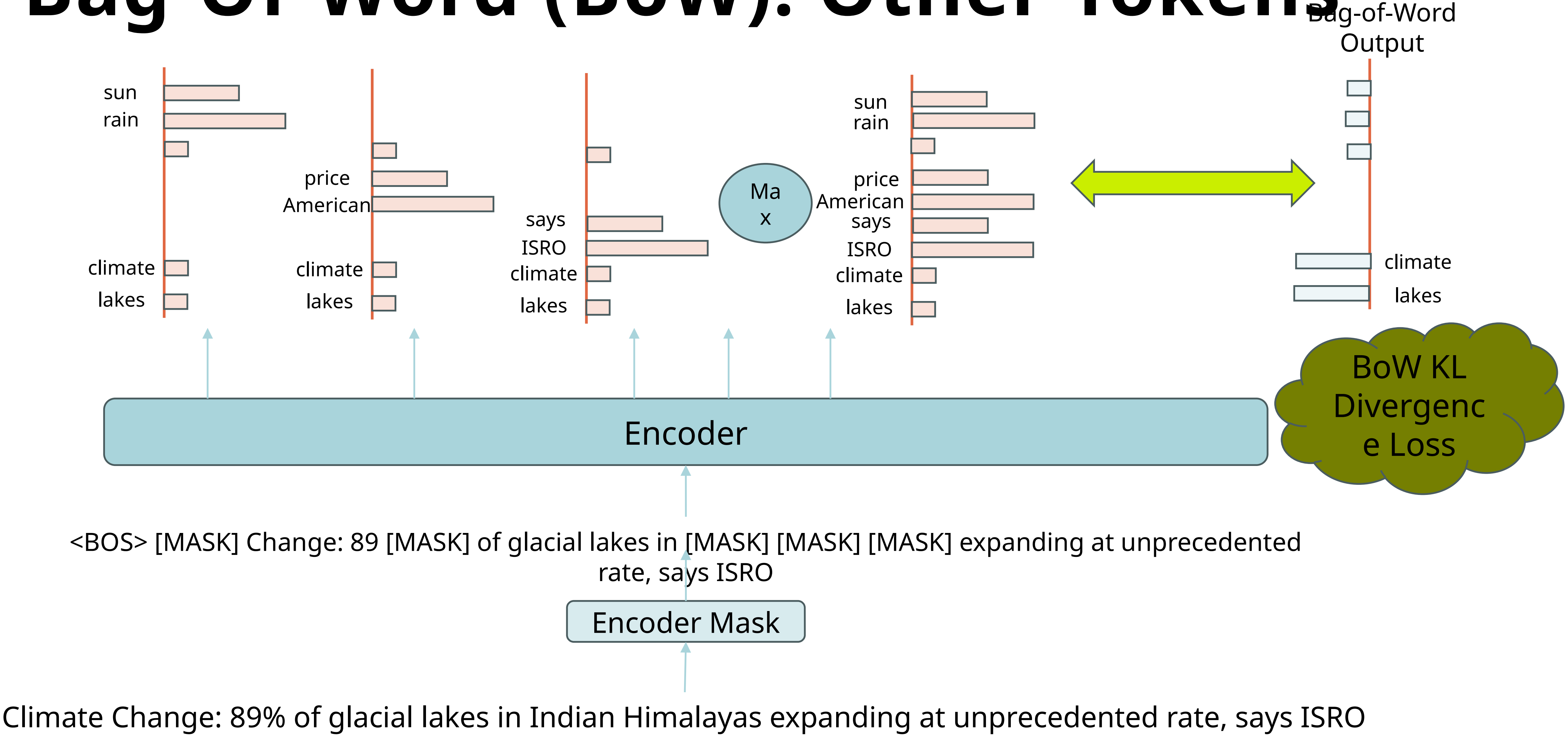
MLM Pre-Training



Decoder Pre-Training



Bag-Of-Word (BoW): Other Tokens



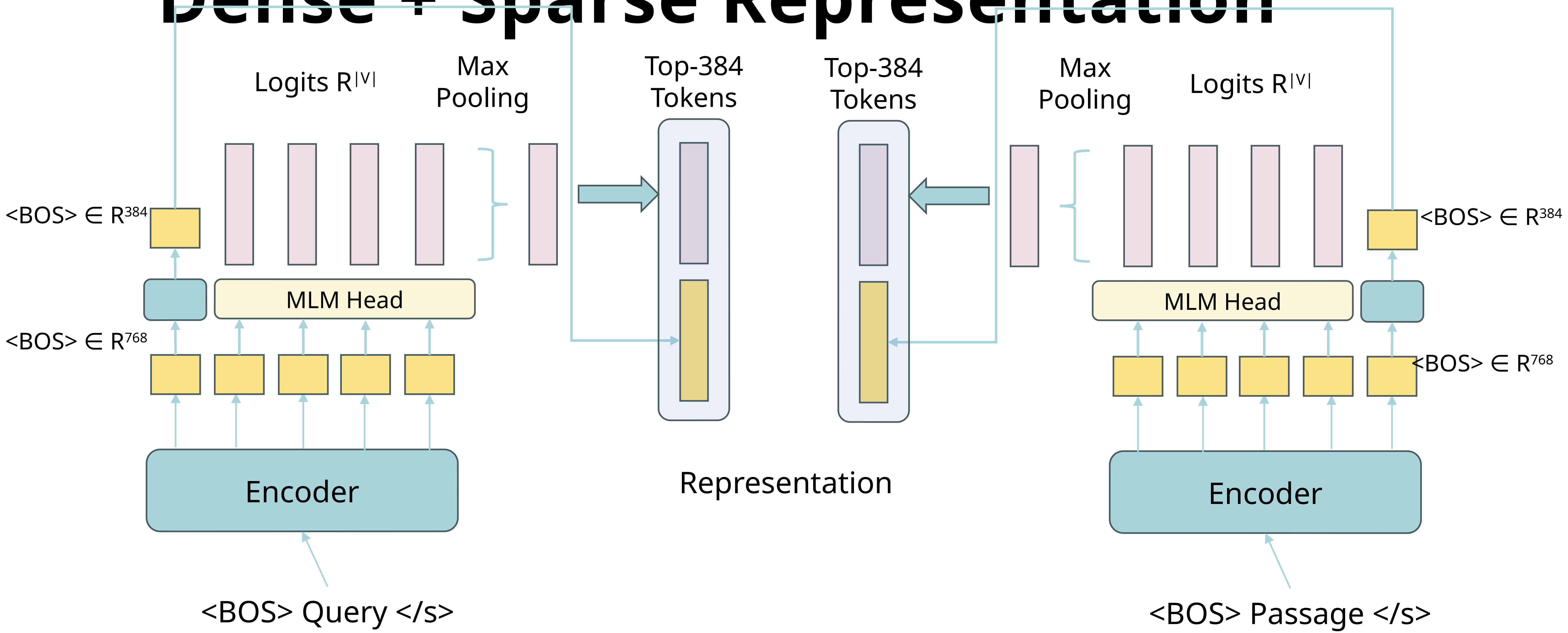
Retro-MAE v2

After pre-training we will have two representations

- <BOS> trained via MLM and enhanced decoding
- Remaining token embeddings trained via BoW reconstruction

FINE-TUNING

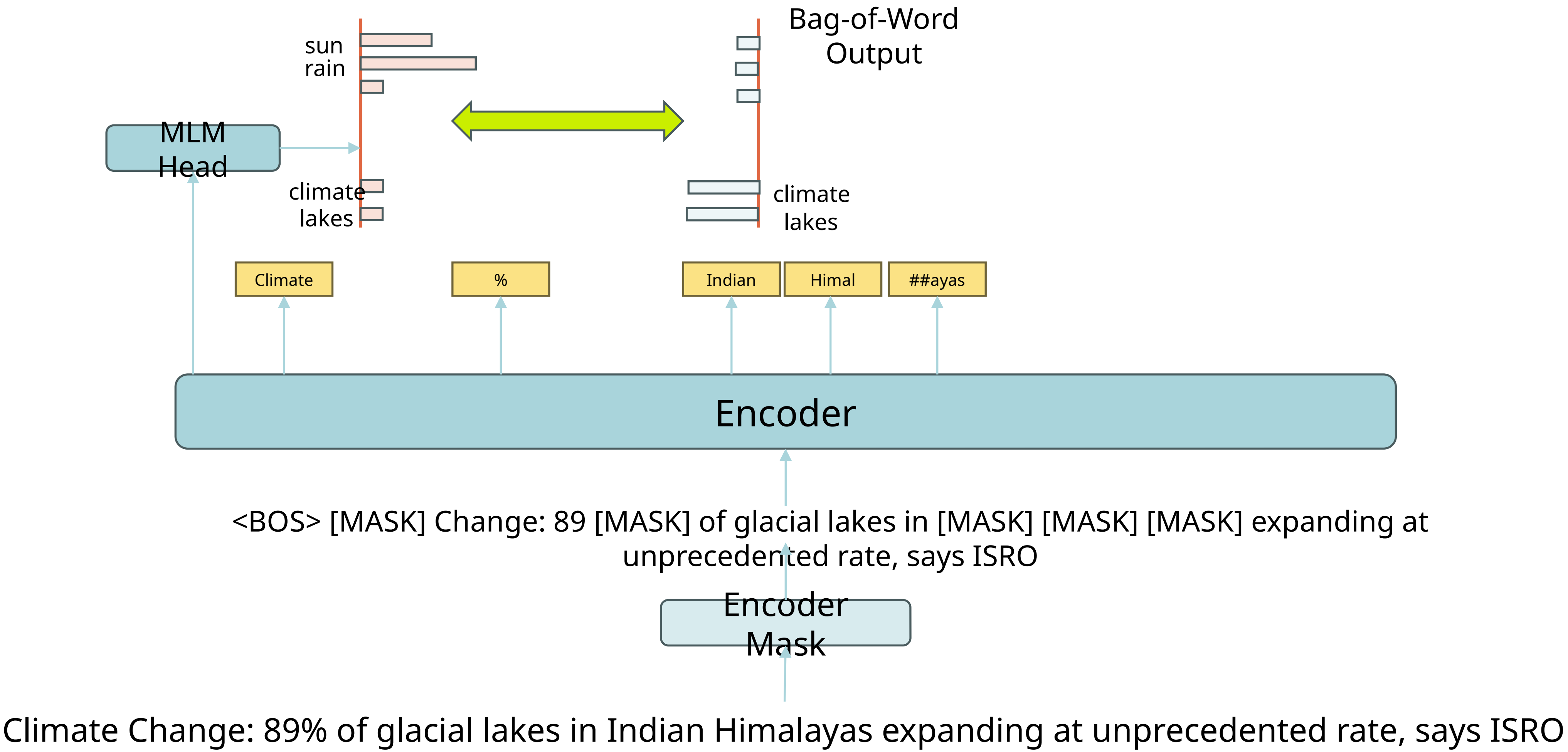
Dense + Sparse Representation



BOW DPR

MA, GUANGYUAN, XING WU, ZIJIA LIN AND SONGLIN HU. “DROP YOUR DECODER: PRE-TRAINING WITH BAG-OF-WORD PREDICTION FOR DENSE PASSAGE RETRIEVAL.” SIGIR 2024

Bag-Of-Word (BoW): CLS



COT MAE

COT-MAE: CONTEXTUAL MASK AUTO-ENCODER FOR DENSE PASSAGE RETRIEVAL. COT-MAE IS A TRANSFORMERS BASED MASK AUTO-ENCODER PRE-TRAINING ARCHITECTURE DESIGNED FOR DENSE PASSAGE RETRIEVAL. (ACCEPTED BY AAAI 2022)

CoT-MAE

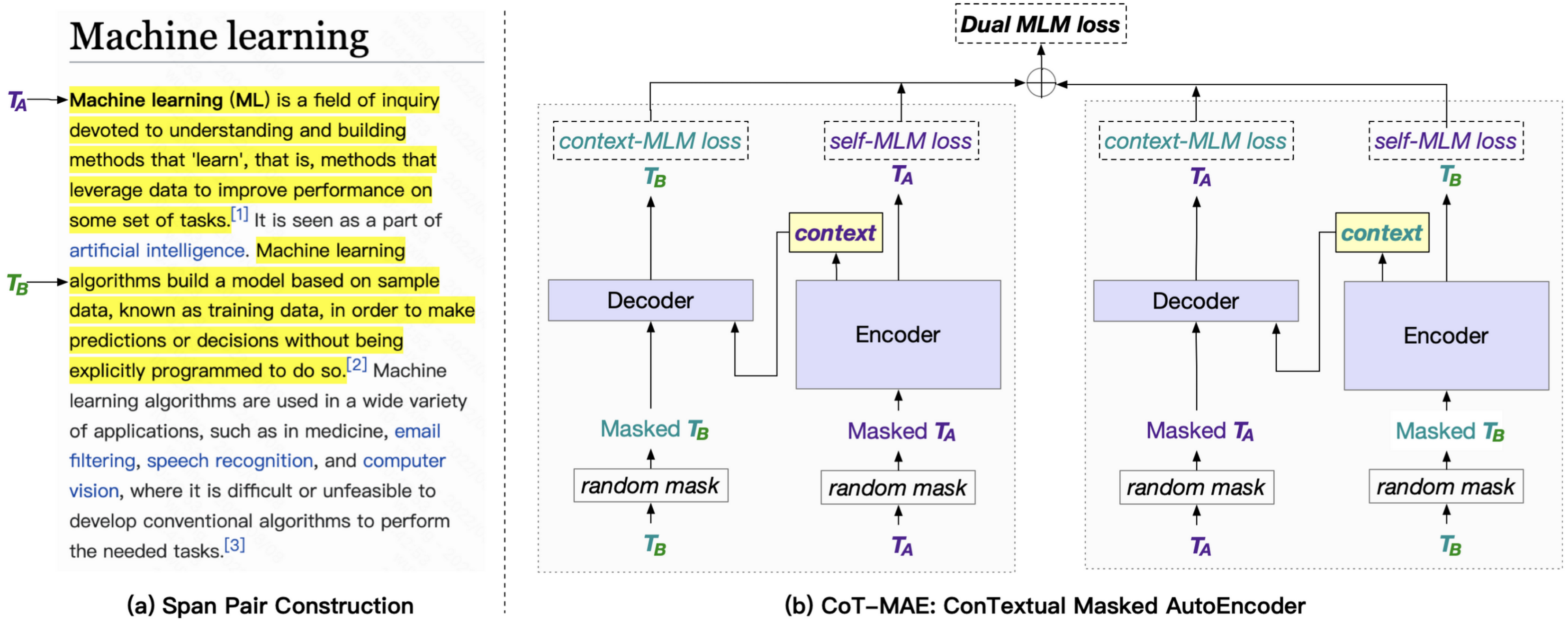
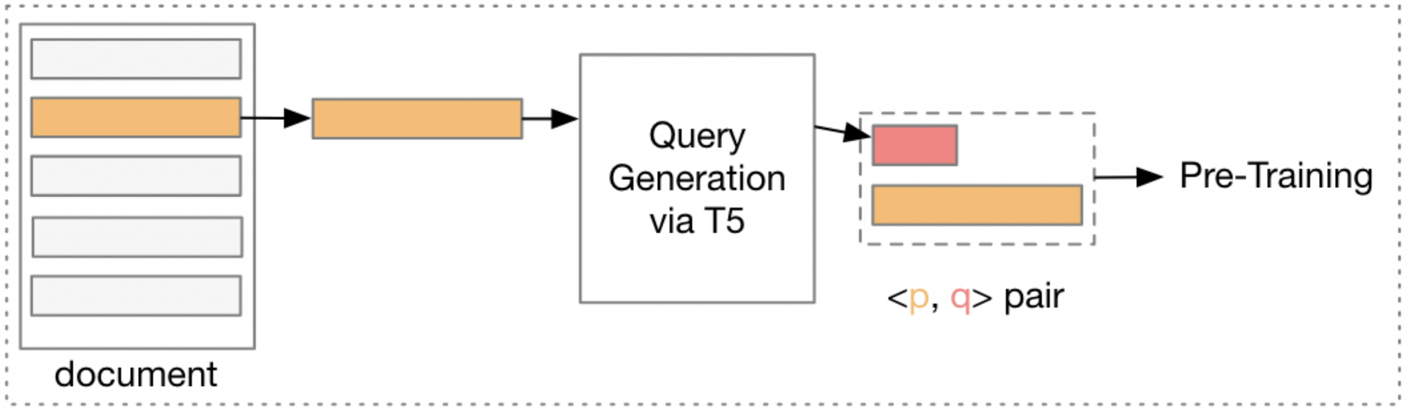


Figure 1: CoT-MAE. (a) The process of span pair construction. We select two neighboring text spans $\mathbf{T_A}$ and $\mathbf{T_B}$ from a document with a sampling strategy to form a span pair. The two spans in a pair are each other's context. (b) The model design for CoT-MAE. We use an asymmetric encoder-decoder structure, with a deep encoder having enough parameters to learn good text representations modeling ability and a shallow decoder to assist the encoder in achieving this goal.

COT-MAE WITH QUERY

XING W, GUANGYUAN MA, WANHUI QIAN, ZIJIA LIN, AND SONGLIN HU. 2023. QUERY-AS-CONTEXT PRE-TRAINING FOR DENSE PASSAGE RETRIEVAL. IN *PROCEEDINGS OF THE 2023 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING*, PAGES 1906–1916, SINGAPORE. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS.

Cot-MAE with Query as Context



(2) Query-as-context Pre-training

COT-MAE V2

WU, XING, GUANGYUAN MA, PENG WANG, MENG LIN, ZIJIA LIN, FUZHENG ZHANG, AND SONGLIN HU. "COT-MAE V2: CONTEXTUAL MASKED AUTO-ENCODER WITH MULTI-VIEW MODELING FOR PASSAGE RETRIEVAL." ARXIV PREPRINT ARXIV:2304.03158 (2023).

CoT-MAE v2

- Multi-view representation learning
  - Dense and Sparse representations
  - Auto-Encoding (MLM) and Auto-Regressive Decoders (CLM)
  - Sparse representation focuses on lexical and dense representation focuses on semantics
- Auto-Encoding Decoder
  - Given sentence representation from encoder and aggressively masked input to the decoder, predict only masked tokens (MLM)
- Auto-Regressive Decoder
  - Given sentence representation from encoder and aggressively masked input to the decoder, reconstruct the original sequence

CoT-MAE v2

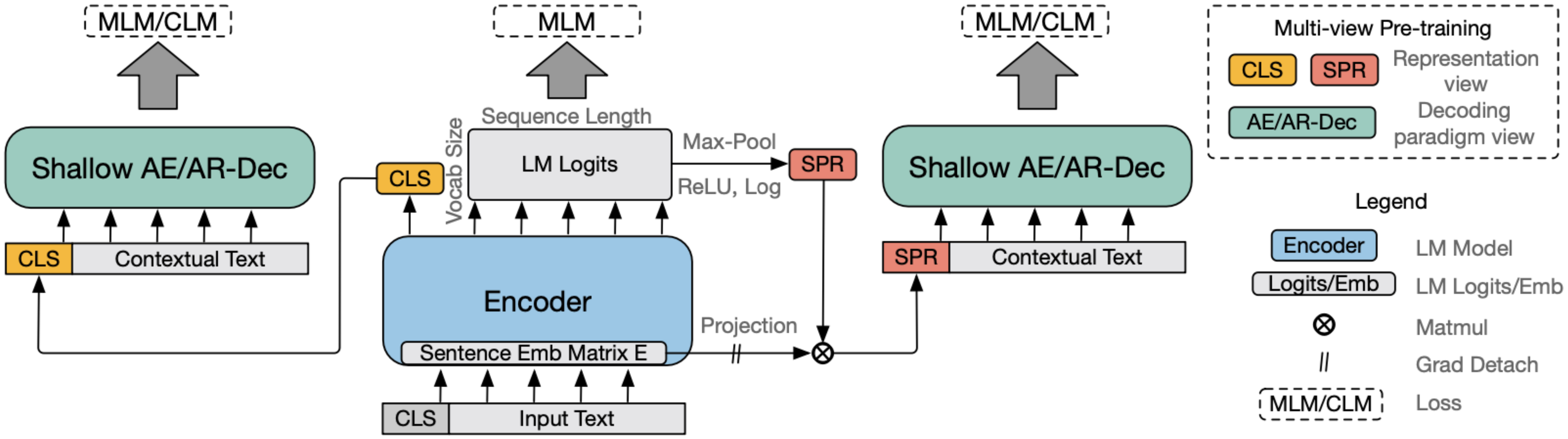


Figure 1: Pre-training designs of CoT-MAE v2. CoT-MAE v2 utilizes both dense (CLS) and sparse (SPR) vectors as multi-view representations. As a multi-view decoding paradigm, Auto-Encoding Decoder (AE-Dec) and Auto-Regressive Decoder (AR-Dec) are integrated into contextual masked auto-encoder pre-training to provide both MLM reconstruction signals and CLM generative signals for representation pre-training.

Evaluation

BEIR (Benchmarking IR)

Split (→)					Train	Dev	Test			Avg. Word Lengths	
Task (↓)	Domain (↓)	Dataset (↓)	Title	Relevancy	#Pairs	#Query	#Query	#Corpus	Avg. D / Q	Query	Document
Passage-Retrieval	Misc.	MS MARCO [42]	✗	Binary	532,761	—-	6,980	8,841,823	1.1	5.96	55.98
Bio-Medical Information Retrieval (IR)	Bio-Medical	TREC-COVID [63]	✓	3-level	—-	—-	50	171,332	493.5	10.60	160.77
	Bio-Medical	NFCorpus [7]	✓	3-level	110,575	324	323	3,633	38.2	3.30	232.26
	Bio-Medical	BioASQ [59]	✓	Binary	32,916	—-	500	14,914,602	4.7	8.05	202.61
Question Answering (QA)	Wikipedia	NQ [32]	✓	Binary	132,803	—-	3,452	2,681,468	1.2	9.16	78.88
	Wikipedia	HotpotQA [74]	✓	Binary	170,000	5,447	7,405	5,233,329	2.0	17.61	46.30
	Finance	FiQA-2018 [41]	✗	Binary	14,166	500	648	57,638	2.6	10.77	132.32
Tweet-Retrieval	Twitter	Signal-1M (RT) [57]	✗	3-level	—-	—-	97	2,866,316	19.6	9.30	13.93
News Retrieval	News	TREC-NEWS [56]	✓	5-level	—-	—-	57	594,977	19.6	11.14	634.79
	News	Robust04 [62]	✗	3-level	—-	—-	249	528,155	69.9	15.27	466.40
Argument Retrieval	Misc.	ArguAna [65]	✓	Binary	—-	—-	1,406	8,674	1.0	192.98	166.80
	Misc.	Touché-2020 [6]	✓	3-level	—-	—-	49	382,545	19.0	6.55	292.37
Duplicate-Question Retrieval	StackEx.	CQADupStack [23]	✓	Binary	—-	—-	13,145	457,199	1.4	8.59	129.09
	Quora	Quora	✗	Binary	—-	5,000	10,000	522,931	1.6	9.53	11.44
Entity-Retrieval	Wikipedia	DBPedia [19]	✓	3-level	—-	67	400	4,635,922	38.2	5.39	49.68
Citation-Prediction	Scientific	SCIDOCS [9]	✓	Binary	—-	—-	1,000	25,657	4.9	9.38	176.19
Fact Checking	Wikipedia	FEVER [58]	✓	Binary	140,085	6,666	6,666	5,416,568	1.2	8.13	84.76
	Wikipedia	Climate-FEVER [13]	✓	Binary	—-	—-	1,535	5,416,593	3.0	20.13	84.76
	Scientific	SciFact [66]	✓	Binary	920	—-	300	5,183	1.1	12.37	213.63

Table 1: Statistics of datasets in BEIR benchmark. Few datasets contain documents without titles. Relevancy indicates the query-document relation: binary (relevant, non-relevant) or graded into sub-levels. Avg. D/Q indicates the average relevant documents per query.

LoTTE (Long-Tail, ToPIC-Stratified Evaluation)

Topic	Question Set	Dev			Test		
		# Questions	# Passages	Subtopics	# Questions	# Passages	Subtopics
Writing	Search Forum	497 2003	277k	ESL, Linguistics, Worldbuilding	1071 2000	200k	English
Recreation	Search Forum	563 2002	263k	Sci-Fi, RPGs, Photography	924 2002	167k	Gaming, Anime, Movies
Science	Search Forum	538 2013	344k	Chemistry, Statistics, Academia	617 2017	1.694M	Math, Physics, Biology
Technology	Search Forum	916 2003	1.276M	Web Apps, Ubuntu, SysAdmin	596 2004	639k	Apple, Android, UNIX, Security
Lifestyle	Search Forum	496 2076	269k	DIY, Music, Bicycles, Car Maintenance	661 2002	119k	Cooking, Sports, Travel

Topic-aligned dev-test pairings

Search queries are from GooAQ linked to StackExchange.
Forum queries are from questions-like StackExchange titles

Datasets

MS MARCO Ranking Test

- The most commonly used IR benchmark
- Adapted from question answering dataset
- More than 500k Bing search queries

TREC

- Text REtrieval Conference (TREC) conducts annual competitions for benchmarking IR systems

Outline

- Introduction
- IR Approaches
- Metrics
- Indic IR

METRICS

Slides Credit: https://www.pinecone.io/learn/offline-evaluation/

Metrics

- How good is our retrieval system?
- Is it able to retrieve the relevant documents given a query?

- **Query:** Impact of climate change on India based on IPCC report
- **Passages:**

Passage 1:
According to the Intergovernmental Panel on Climate Change (IPCC), India is among the countries that face the highest risk from climate change's impact, despite contributing minimally to global warming in the past century. The IPCC's 2022 report on climate change impacts and risks to ecosystems and human systems highlights so..

Passage 2:
India may face catastrophic impacts due to global warming, IPCC reports warn. Temperature rise, sea level increase, catastrophic impacts on the lives and livelihoods of people are some of the big challenges for India ...

Passage 3:
The occurrence of extreme hot events is likely to increase in Indiana, while the occurrence of extreme cold events is likely to decrease. The occurrence of conditions that spawn severe thunderstorms is likely to increase in Indiana...

Passage 4:
Climate change is expected to have major health impacts in India- increasing malnutrition and related health disorders such as child stunting - with the poor likely to be affected most severely..

Passage 5:
As global temperatures rise, we will continue to see India's poorest suffer the most as **climate change destroys livelihoods and washes away…**

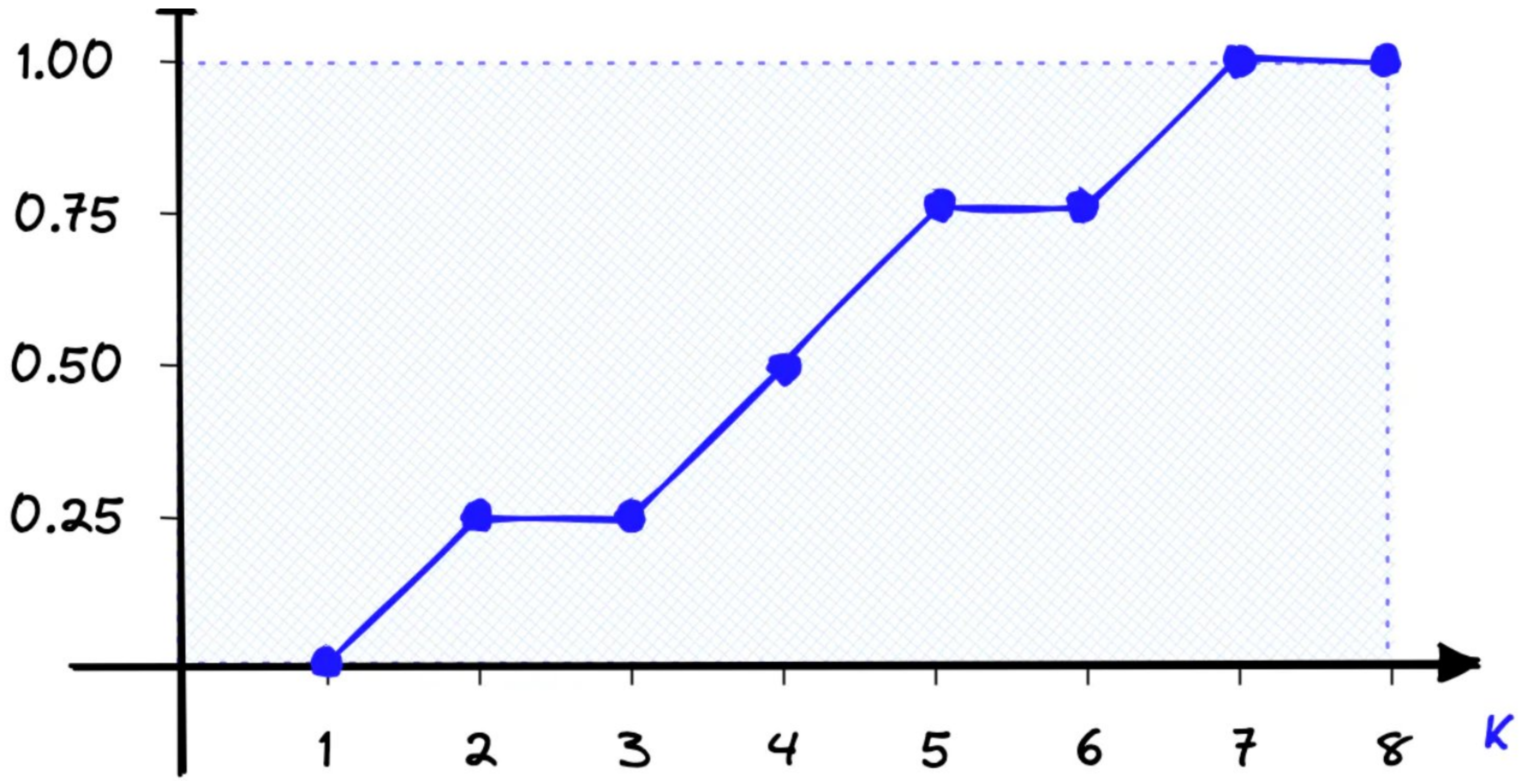
Example

Recall@k

- Recall@K measure how many relevant items were returned out of all the relevant items
- Consider Recall@3
- There are only 2 relevant item
- **Scenario 1:**
  - Returned Items = {P1, P3, P4}
  - Recall@3 = $\frac{1}{2}$
- **Scenario 2:**
  - Returned Items = {P1, P2, P4}
  - Recall@3 = $\frac{2}{2}$
- **Scenario 3:**
  - Returned Items = {P3, P5, P1}
  - Recall@3 = $\frac{1}{2}$

Recall@k

Source: https://www.pinecone.io/learn/offline-evaluation/



Recall@k

Summary

- Easy to understand and interpret
- A perfect score indicates all relevant items are returned
- A value of zero indicates no relevant items are returned
- It is order unaware
- A system which returns a relevant item at position 1 is given the same score as another system which returns a relevant item at position 10 for Recall@10

Mean Reciprocal Rank (MRR)

- Unlike Recall@K, MRR is an **order-aware metric**
- $MRR = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{rank_q}$
- Q is the total number of queries in your test set
- $rank_q$ is the rank of the first *relevant result* for query q

MRR

- Assume we have 3 different queries in our test set
- The relevant document will be in bold
- **Query 1:**
  - Returned Items = {P1, **P3**, P4}
  - $rank_q = \frac{1}{2} = 0.5$
- **Query 2:**
  - Returned Items = {**P1**, P2, P4}
  - $rank_q = \frac{1}{1} = 1$
- **Query 3:**
  - Returned Items = {P3, P5, **P1**}
  - $rank_q = \frac{1}{3} = 0.33$
- MRR = $\frac{0.5+1+0.33}{3} = 0.61$

MRR

Summary

- Order-aware makes it more relevant for use-cases where the ranking of the relevant result is important
- We consider, rank of the first relevant item only

Summary

- Brief overview of IR methods
- Classical Approaches as well as encoder-based approaches
- Various metrics for evaluation

References

- Introduction to Information Retrieval. C.D. Manning, P. Raghavan, H. Schütze. Cambridge UP, 2008.
- Information Retrieval: An Introduction. Dr. Grace Hui Yang, InfoSense, 2019
- Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. "Dense Passage Retrieval for Open-Domain Question Answering." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

References

- Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., ... & Wei, F. (2023, July). SimLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2244-2258).
- Chang, Wei-Cheng, X. Yu Felix, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. "Pre-training Tasks for Embedding-based Large-scale Retrieval." In International Conference on Learning Representations.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint arXiv:2104.08663
- Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20). Association for Computing Machinery, New York, NY, USA, 39–48.

REFERENCES

- Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4):333–389

- Keshav Santhanam, Jon Saad-Falcon, Martin Franz, Omar Khattab, Avi Sil, Radu Florian, Md Arafat Sultan, Salim Roukos, Matei Zaharia, and Christopher Potts. 2023. Moving Beyond Downstream Task Accuracy for Information Retrieval Benchmarking. In Findings of the Association for Computational Linguistics: ACL 2023, pages 11613–11628, Toronto, Canada. Association for Computational Linguistics.

References

- Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Zheng Liu, Shitao Xiao, Yingxia Shao, and Zhao Cao. 2023. RetroMAE-2: Duplex Masked Auto-Encoder For Pre-Training Retrieval-Oriented Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2635–2648, Toronto, Canada. Association for Computational Linguistics.

References

- **CoT-MAE: ConTextual Mask Auto-Encoder for Dense Passage Retrieval. CoT-MAE is a transformers based Mask Auto-Encoder pre-training architecture designed for Dense Passage Retrieval. (Accepted by AAAI 2022)**

- **Ma, Guangyuan, Xing Wu, Zijia Lin and Songlin Hu. “Drop your Decoder: Pre-training with Bag-of-Word Prediction for Dense Passage Retrieval.” SigIR 2024**

- Xing W, Guangyuan Ma, Wanhui Qian, Zijia Lin, and Songlin Hu. 2023. Query-as-context Pre-training for Dense Passage Retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1906–1916, Singapore. Association for Computational Linguistics.

References

- Wu, Xing, Guangyuan Ma, Peng Wang, Meng Lin, Zijia Lin, Fuzheng Zhang, and Songlin Hu. "Cot-mae v2: contextual masked auto-encoder with multi-view modeling for passage retrieval." arXiv preprint arXiv:2304.03158 (2023).