INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

IIIT
Hyderabad

DENSE RETRIEVAL FOR LOCAL/LOW-RESOURCE LANGUAGES

October 2023

COURSE CODE: INFORMATION RETRIEVAL & EXTRACTION – CS4.406.M23

Advisor:
Prof. Rahul Mishra

Mentor:
Ankita Maity

Team Number:
4

Representatives:
Aadesh Ingle - 2022202017
Akhilesh Giriboyina - 2022202022
Samyak Jain – 2022201048

Academic year:
2023-2024

1. **Project Overview**:

Problem Statement: The problem statement involves developing an effective and efficient dense retrieval system for local or low-resource languages. Dense retrieval focuses on retrieving relevant passages or documents from a collection based on dense vector representations, typically obtained from pre-trained language models. In the context of local or low-resource languages, the challenge is to adapt existing retrieval techniques and models to manage limited data and linguistic resources while maintaining high retrieval performance.

Understanding of the Problem Statement: The problem entails addressing several key aspects:

- Creating high-quality dense embeddings for passages/documents in the low-resource language.
- Designing an efficient indexing and retrieval mechanism for these embeddings.
- Adapting or fine-tuning pre-trained models to handle data scarcity of the target language.
- Evaluating the system's effectiveness in retrieving relevant information in the low-resource language.

2. **Data Overview**:

Here are some of the datasets which we have reviewed:

- Wikipedia Dumps

  A complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML. A number of raw database tables in SQL form are also available. These snapshots are provided at the very least monthly and usually twice a month.

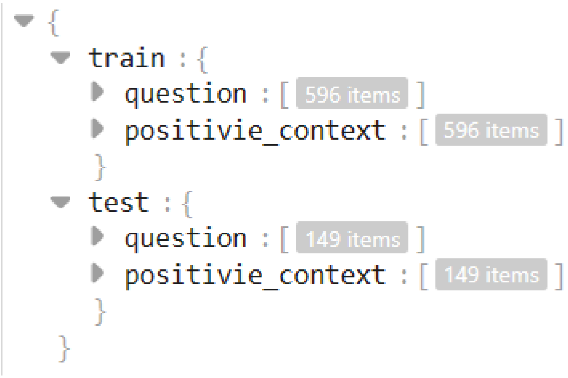
- ChAII(trans) and ChAII

  The dataset covers Marathi, Hindi and Tamil, collected without the use of translation. It provides a realistic information-seeking task with questions written by native-speaking expert data annotators.

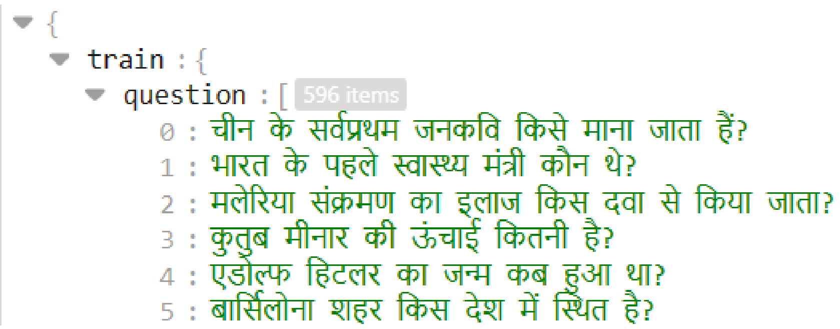
- xQUAD
  XQuAD (Cross-lingual Question Answering Dataset) is a benchmark dataset for evaluating cross-lingual question answering performance. The dataset consists of a subset of 240 paragraphs and 1190 question-answer pairs from the development set of SQuAD v1.1 (Rajpurkar et al., 2016), together with their professional translations into ten languages: Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi. Consequently, the dataset is entirely parallel across 11 languages.

- The Indic NLP Catalog
  A Collaborative Catalog of Resources for Indic Language NLP

For the interim submission, we have decided to go with the **ChAII hindi dataset**. We have taken the questions, & positive contexts from the dataset. We have customized this dataset to our own needs. Following is the customized format for our dataset, which is different from the original dataset.

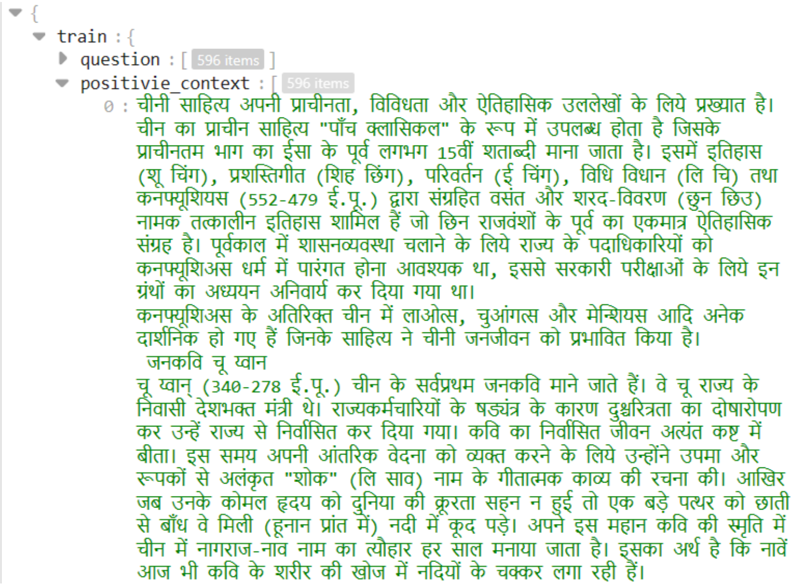


```
{
  train : {
    question : [ 596 items ]
    positivie_context : [ 596 items ]
  }
  test : {
    question : [ 149 items ]
    positivie_context : [ 149 items ]
  }
}
```



```
{
  train : {
    question : [ 596 items
      0 : चीन के सर्वप्रथम जनकवि किसे माना जाता हैं?
      1 : भारत के पहले स्वास्थ्य मंत्री कौन थे?
      2 : मलेरिया संक्रमण का इलाज किस दवा से किया जाता?
      3 : कुतुब मीनार की ऊंचाई कितनी है?
      4 : एडोल्फ हिटलर का जन्म कब हुआ था?
      5 : बार्सिलोना शहर किस देश में स्थित है?
```

Similar format is followed for test set.

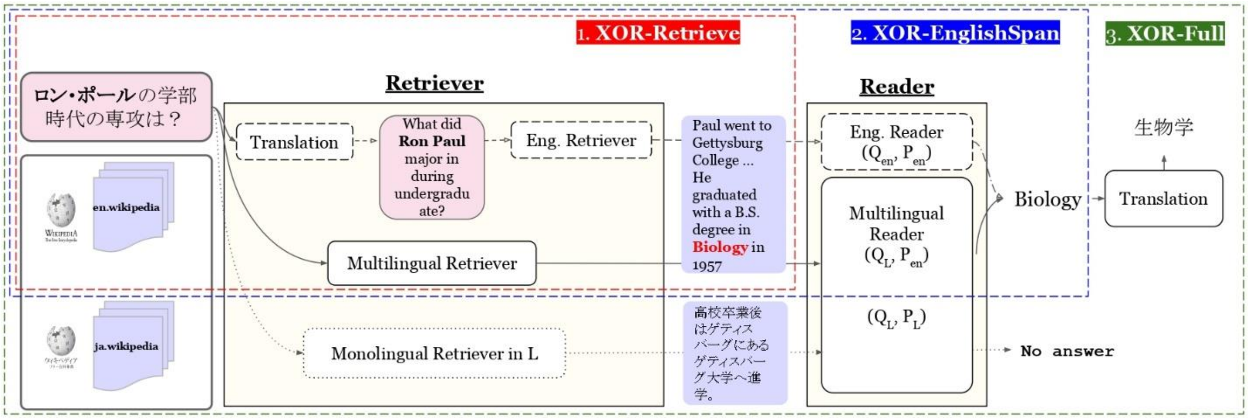


```
{
  train : {
    question : [ 596 items ]
    positivie_context : [ 596 items
      0 : चीनी साहित्य अपनी प्राचीनता, विविधता और ऐतिहासिक उललेखों के लिये प्रख्यात है।
          चीन का प्राचीन साहित्य "पाँच क्लासिकल" के रूप में उपलब्ध होता है जिसके
          प्राचीनतम भाग का ईसा के पूर्व लगभग 15वीं शताब्दी माना जाता है। इसमें इतिहास
          (शू चिंग), प्रशस्तिगीत (शिह छिंग), परिवर्तन (ई चिंग), विधि विधान (लि चि) तथा
          कनफ्यूशियस (552-479 ई.पू.) द्वारा संग्रहित वसंत और शरद-विवरण (छुन छिउ)
          नामक तत्कालीन इतिहास शामिल हैं जो छिन राजवंशों के पूर्व का एकमात्र ऐतिहासिक
          संग्रह है। पूर्वकाल में शासनव्यवस्था चलाने के लिये राज्य के पदाधिकारियों को
          कनफ्यूशिअस धर्म में पारंगत होना आवश्यक था, इससे सरकारी परीक्षाओं के लिये इन
          ग्रंथों का अध्ययन अनिवार्य कर दिया गया था।
          कनफ्यूशिअस के अतिरिक्त चीन में लाओत्स, चुआंगत्स और मेन्शियस आदि अनेक
          दार्शनिक हो गए हैं जिनके साहित्य ने चीनी जनजीवन को प्रभावित किया है।
           जनकवि चू य्वान
          चू य्वान् (340-278 ई.पू.) चीन के सर्वप्रथम जनकवि माने जाते हैं। वे चू राज्य के
          निवासी देशभक्त मंत्री थे। राज्यकर्मचारियों के षड्यंत्र के कारण दुश्चरित्रता का दोषारोपण
          कर उन्हें राज्य से निर्वासित कर दिया गया। कवि का निर्वासित जीवन अत्यंत कष्ट में
          बीता। इस समय अपनी आंतरिक वेदना को व्यक्त करने के लिये उन्होंने उपमा और
          रूपकों से अलंकृत "शोक" (लि साव) नाम के गीतात्मक काव्य की रचना की। आखिर
          जब उनके कोमल हृदय को दुनिया की क्रूरता सहन न हुई तो एक बड़े पत्थर को छाती
          से बाँध वे मिली (हूनान प्रांत में) नदी में कूद पड़े। अपने इस महान कवि की स्मृति में
          चीन में नागराज-नाव नाम का त्यौहार हर साल मनाया जाता है। इसका अर्थ है कि नावें
          आज भी कवि के शरीर की खोज में नदियों के चक्कर लगा रही हैं।
```

3. Our approach using DPR:

We have reviewed various papers and online resources for our project topic. We have tweaked the DPR paper along with some modifications as we do not have access to the Institute's ADA cluster.

Low-Resource Dense Retrieval for Open-Domain Question Answering


1. XOR-Retrieve
2. XOR-EnglishSpan
3. XOR-Full
Retriever
Reader
ロン・ポールの学部時代の専攻は？
en.wikipedia
ja.wikipedia
Translation
What did Ron Paul major in during undergraduate?
Eng. Retriever
Multilingual Retriever
Monolingual Retriever in L
Paul went to Gettysburg College ... He graduated with a B.S. degree in Biology in 1957
高校卒業後はゲティスバーグにあるゲティスバーグ大学へ進学。
Eng. Reader (Q_{en}, P_{en})
Multilingual Reader (Q_L, P_{en})
(Q_L, P_L)
Biology
Translation
生物学
No answer

- It provides a comprehensive overview of the state-of-the-art methods for low-resource dense retrieval (DR) for both the document collection and a set of question-answer pairs to learn a dense retriever. This setting is the least challenging, as the retriever can learn from the question-answer pairs to directly predict the answer to the question.

Major Motivation from the above paper: We have decided to go for the documents-only dataset that makes the question generation model task the most important for the Retrieval model.

Our proposed implementation:

Given a collection of M text passages, the goal of our dense passage retriever (DPR) is to index all the passages in a low-dimensional and continuous space such that it can efficiently retrieve the top k passages relevant to the input question for the reader at run-time.

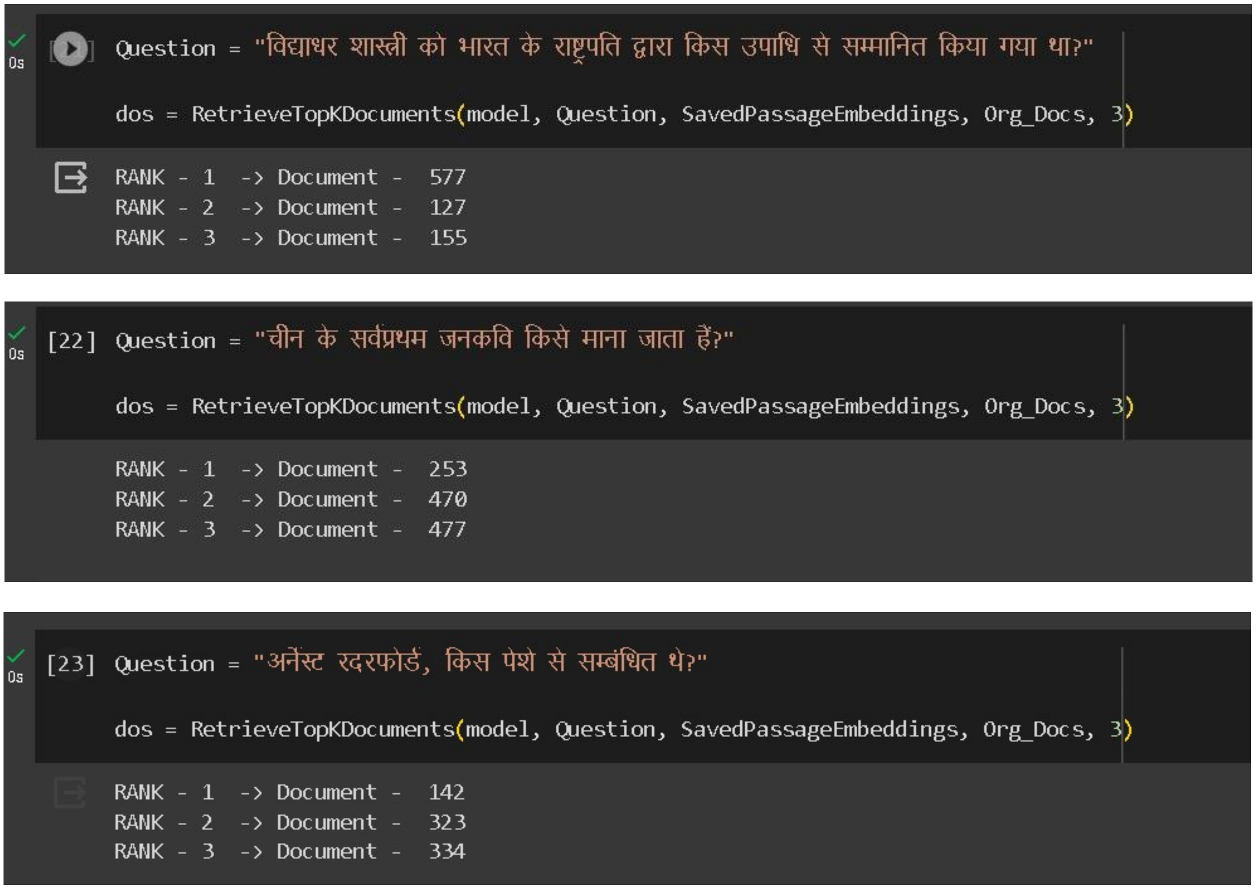
- Our dense passage retriever (DPR) uses a **dense encoder EP (·),** which maps any text passage to a d-dimensional real-valued vector and builds an index for all the M passages that we will use for retrieval.

- At run-time, DPR applies a different **encoder EQ (·)** that maps the input question to a d-dimensional vector and retrieves k passages of which vectors are the closest to the question vector.
- We define the similarity between the question and the passage using the dot product of their vectors: $\mathrm{sim}(q, p) = E_Q(q)^{\intercal} E_P(p)$
- **ENCODER:** Although, in principle, the question and passage encoders can be implemented by any neural network, in this work, we use two independent BERT networks.
- **INFERENCING:** Given a question q at run-time, we derive its embedding $v_q = E_Q(q)$ and retrieve the top k passages with embeddings closest to v_q.
- **TRAINING:**
  - The goal is to create a vector space such that relevant pairs of questions and passages will have a smaller distance (i.e., higher similarity) than the irrelevant ones by learning a better embedding function.
- **CUSTOM LOSS FUNCTION:**
  - Let $\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^m$ be the training data that consists of m instances. Each instance contains one question q_i and one relevant (positive) passage p_i^+ along with n irrelevant (negative) passages $p_{i,j}^-$. We optimize the loss function as the negative log-likelihood of the positive passage: $$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\mathrm{sim}(q_i, p_{i,j}^-)}}$$
- **POSITIVE AND NEGATIVE PASSAGES:**
  - For retrieval problems, it is often the case that positive examples are available explicitly, while negative examples need to be selected from an extremely large pool.
  - For instance, passages relevant to a question may be given in a QA dataset or can be found using the answer. All other passages in the collection, while not specified explicitly, can be viewed as irrelevant by default.
  - We have decided to do in **BATCH NEGATIVES**.

- **IN BATCH NEGATIVES:**
  - Assume that we have B questions in a mini-batch, and each one is associated with a relevant passage. Let **Q** and **P** be the (B×d) matrix of question and passage embeddings in a batch of size B.

- $S = QP^T$ is a (B ×B) matrix of similarity scores, where each row of which corresponds to a question paired with B passages.
- In this way, we **reuse computation** and effectively train on B^2 (q_i, p_j) question/passage pairs in each batch.
- Any (q_i, p_j) pair is a positive example when i == j and negative otherwise.
- This creates B training instances in each batch, where there are **B – 1 negative passages** for each question.

Our Model Output:



```
Question = "विद्याधर शास्त्री को भारत के राष्ट्रपति द्वारा किस उपाधि से सम्मानित किया गया था?"

dos = RetrieveTopKDocuments(model, Question, SavedPassageEmbeddings, Org_Docs, 3)
```

```
RANK - 1  -> Document -  577
RANK - 2  -> Document -  127
RANK - 3  -> Document -  155
```

```
[22] Question = "चीन के सर्वप्रथम जनकवि किसे माना जाता हैं?"

     dos = RetrieveTopKDocuments(model, Question, SavedPassageEmbeddings, Org_Docs, 3)
```

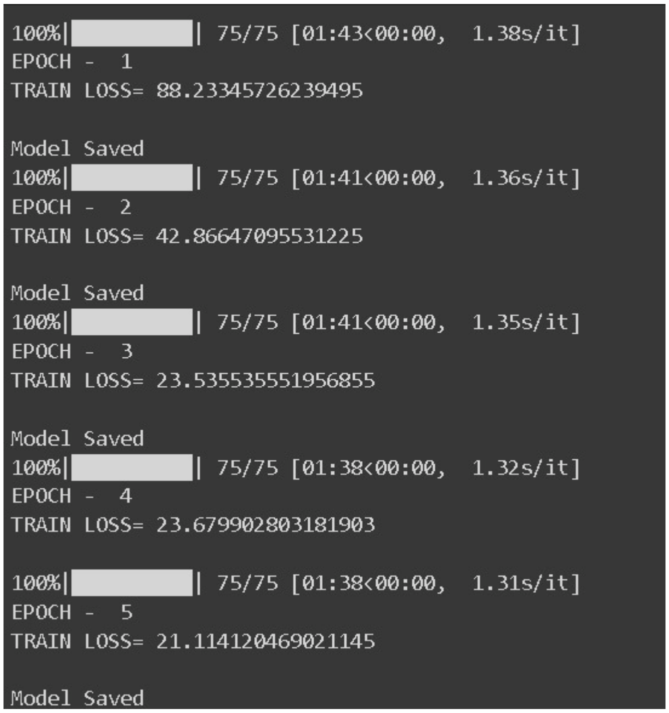
```
RANK - 1  -> Document -  253
RANK - 2  -> Document -  470
RANK - 3  -> Document -  477
```

```
[23] Question = "अर्नेस्ट रदरफोर्ड, किस पेशे से सम्बंधित थे?"

     dos = RetrieveTopKDocuments(model, Question, SavedPassageEmbeddings, Org_Docs, 3)
```

```
RANK - 1  -> Document -  142
RANK - 2  -> Document -  323
RANK - 3  -> Document -  334
```

Figure 1: These questions are taken from test dataset which were not included as part of our training dataset.



```
100%|██████████| 75/75 [01:43<00:00,  1.38s/it]
EPOCH -  1
TRAIN LOSS= 88.23345726239495

Model Saved
100%|██████████| 75/75 [01:41<00:00,  1.36s/it]
EPOCH -  2
TRAIN LOSS= 42.86647095531225

Model Saved
100%|██████████| 75/75 [01:41<00:00,  1.35s/it]
EPOCH -  3
TRAIN LOSS= 23.535535551956855

Model Saved
100%|██████████| 75/75 [01:38<00:00,  1.32s/it]
EPOCH -  4
TRAIN LOSS= 23.679902803181903

100%|██████████| 75/75 [01:38<00:00,  1.31s/it]
EPOCH -  5
TRAIN LOSS= 21.114120469021145

Model Saved
```

Figure 2 training process

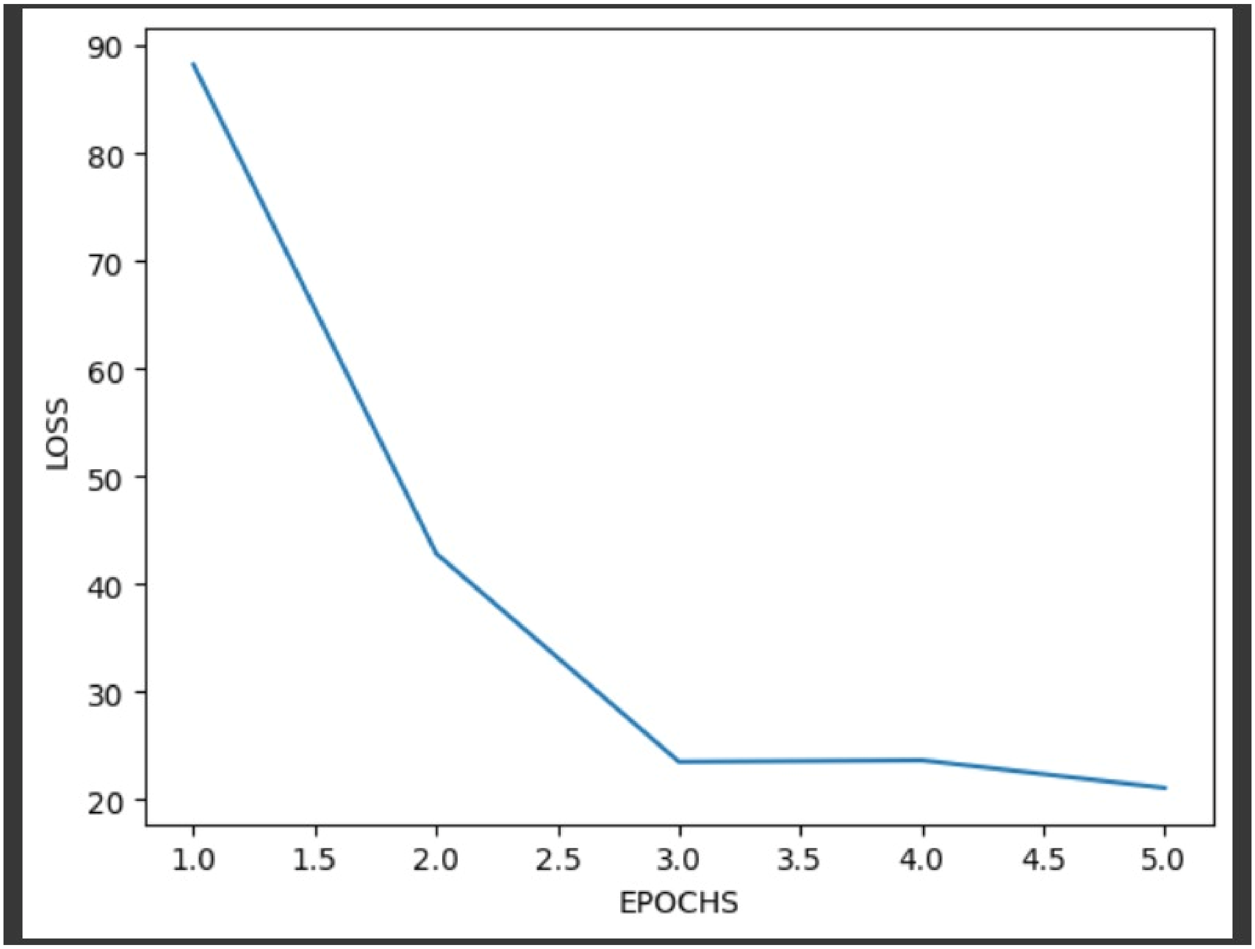


Figure 3 variation is training loss over epochs

4. Code Structure:

a. DataLoader

We are sending question and it’s corresponding context pairs in each batch. We have used in batch negative passages as mentioned earlier to speed up the process and save compute resources.

b. Encoder

We will be using class for generating the [CLS] embeddings of question/ query and paragraph, which we will be using further.

c. Bidirectional Encoder:

We will use this class as the model class that we will train. Here, we are using the above Encoder to get the [CLS] embeddings of both question/ query and paragraph.

d. Train Bidirectional Encoder:

We are running the train loop here and finetuning the BERT model to make embeddings of question and positive paragraphs as close as possible and vice versa for question and negative paragraphs.

Custom Loss Function: We are using NLL Loss(Negative Log-Likelihood), which is generally used in tasks like classification.

e. LoadOrTrainModel:

Finally, in this loop, we are running training the above class for a certain no of epochs, saving the state of the model if the loss is minimum at that epoch, and finally plotting the change in loss over the epoch.

5. Implementation Plan:

Task No.	Task	Estimated Completion Date
1	Literature Review	07/09/2023(Done) ✓
2	Dataset Finalization	13/09/2023 - 18/09/2023 (Done) ✓
3	Question Generation Model	12/10/2023 - 18/10/2023 (ADA access got delayed)
4	Training Document Retrieval Model	10/11/2023 - 15/11/2023 (Done) ✓
5	Evaluation Metrics & Benchmarks	15/11/2023 - 18/11/2023
6	Documentation and Reporting	17/11/2023 - 19/11/2023

**This rough schedule is liable to change or be delayed or not followed based on requirements and further research and input.*

6. Final deliverables:

- The final deliverables consist of a fully functional retrieval system for low-resource languages, a report, comprehensive evaluation results, and a presentation or demo showcasing the system's capabilities.

NOTE:

Initially, we tried to generate augmented data for low-resource language using the mT5 model. But we got stuck in this approach because of limited resources of computing and hard disk for loading mT5 and mC4 corpus.
As suggested by Prof Rahul Mishra, we tried to get ADA access, but the same has still not been granted to us. Hence, we could not successfully proceed with this approach. After the interim, we are planning to re-implement this approach based on the availability of resources using ADA if the same is granted to us.

Google Drive Link:
https://drive.google.com/drive/folders/11PJv7jlOru8E7exs4Pl_yG1i_JIL7yIk?usp=share_link

References:

- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *ArXiv*. /abs/2004.04906
- Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2020). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. *ArXiv*. /abs/2010.12309
- Hambarde, K. A., & Proenca, H. (2023). Information Retrieval: Recent Advances and Beyond. *ArXiv*. /abs/2301.08801
- Kulshreshtha, D., Belfer, R., Serban, I. V., & Reddy, S. (2021). Back-Training excels Self-Training at Unsupervised Domain Adaptation of Question Generation and Passage Retrieval. *ArXiv*. /abs/2104.08801
- Shen, X., Vakulenko, S., Del Tredici, M., Barlacchi, G., Byrne, B., & De Gispert, A. (2022). Low-Resource Dense Retrieval for Open-Domain Question Answering: A Comprehensive Survey. *ArXiv*. /abs/2208.03197
- Wang, K., Thakur, N., Reimers, N., & Gurevych, I. (2021). GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. *ArXiv*. /abs/2112.07577
- Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *ArXiv*. /abs/2104.08663
- Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *ArXiv*. /abs/2004.12832
- Lu, W., Jiao, J., & Zhang, R. (2020). TwinBERT: Distilling Knowledge to Twin-Structured BERT Models for Efficient Retrieval. *ArXiv*. /abs/2002.06275