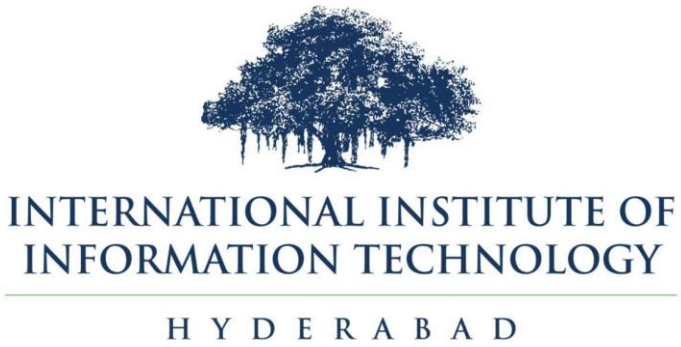
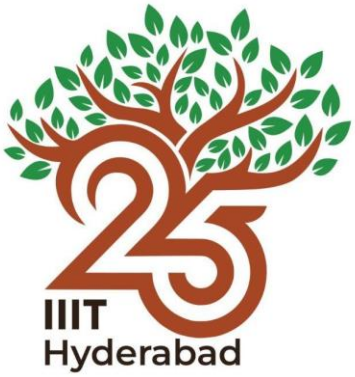




DENSE RETRIEVAL FOR LOCAL/LOW-RESOURCE LANGUAGES

September 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

3. Literature Review:

We have reviewed various papers and online resources for our project topic. Here are some of the most important resources:

1) Low-Resource Dense Retrieval for Open-Domain Question Answering

It provides a comprehensive overview of the state-of-the-art methods for low-resource dense retrieval (DR) for open-domain question answering (QA).

The paper divides the low-resource DR techniques into three main categories based on the dataset available:

- **Documents**: Techniques that only use the document collection to learn a dense retriever. This is the most challenging setting, as the retriever must learn to capture the semantic similarity between questions and documents without supervision.
- **Documents and questions**: Techniques use both the document collection and a set of question-document pairs to learn a dense retriever. This setting is less challenging than the previous one, as the retriever can learn from the question-document pairs to better understand the semantic similarity between questions and documents.
- **Documents and question-answer pairs**: Techniques use 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.

2) Back-Training excels Self-Training at Unsupervised Domain Adaptation of Question Generation and Passage Retrieval

This proposes a new method called back-training for unsupervised domain adaptation of question generation and passage retrieval tasks.

- **Unsupervised domain adaptation** is the task of training a model on data from a source domain and then adapting it to perform well on data from a target domain, even though the two domains have different distributions. This is a challenging task because the model can easily overfit the source domain data and not generalize well to the target domain data.
- **Self-training** is a popular method for unsupervised domain adaptation. It works by first training a model on the source domain data. Then, the model is used to generate predictions on the target domain data. These predictions are then added to the training data, and the model is retrained. This process is repeated until the model converges.
- The **back-training method** works differently. It first trains a model on the source domain data. Then, it generates synthetic target domain data by aligning noisy inputs with natural outputs. The synthetic data is then added to the training data, and the model is retrained.
- The authors of the paper found that back-training outperforms self-training on the task of question generation and passage retrieval. They also found that back-training is more robust to noise in the synthetic data.
- In addition to proposing back-training, the paper also proposes a new **consistency filte**r to remove low-quality synthetic data before training. The consistency filter works by comparing the outputs of the model on the synthetic data to the outputs of the model on the source domain data. If the outputs are not consistent, then the synthetic data is removed.

Major Motivation from the above paper: We might go with the above approach for the Question Generation Model needed for our architecture of the retrieval system.

3) A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios

In the context of the prevalent use of deep neural networks and large language models in natural language applications, this paper explores methods to enhance performance in low-resource settings. It covers data augmentation, cross-lingual projections, and transfer learning approaches, emphasizing their distinct requirements and outlining future research directions.

- The author discussed the application of **data augmentation** techniques in natural language processing (NLP) to generate new instances from existing ones without changing their labels. These methods encompass token-level replacements, such as synonyms or similar entities, and sentence-level operations, like dependency tree manipulation or back-translation. Adversarial methods, which perturb input data without altering meaning, are also explored. The challenges include the need for task-specific understanding and the absence of a unified framework for cross-tasking and cross-language augmentation. Despite this, data augmentation holds promise, particularly in low-resource settings, where it complements pretraining in transformer models and leverages linguistic expertise.
- The author further proposed the use of **cross-lingual projections** as a method to obtain labelled data for low-resource languages in specific NLP tasks. This approach involves training a task-specific classifier in a high-resource language and aligning unlabeled low-resource data to its high-resource counterpart using parallel corpora. Labels from the high-resource text are projected back to the low-resource language based on token alignment. This technique has been applied to various tasks, such as POS tagging and parsing. However, challenges include the need for auxiliary data with labels in a high-resource language and potential difficulties in machine translation for specific low-resource languages. Alternative approaches based on word translations, bilingual dictionaries, and task-specific seed words have also been proposed to address these limitations.

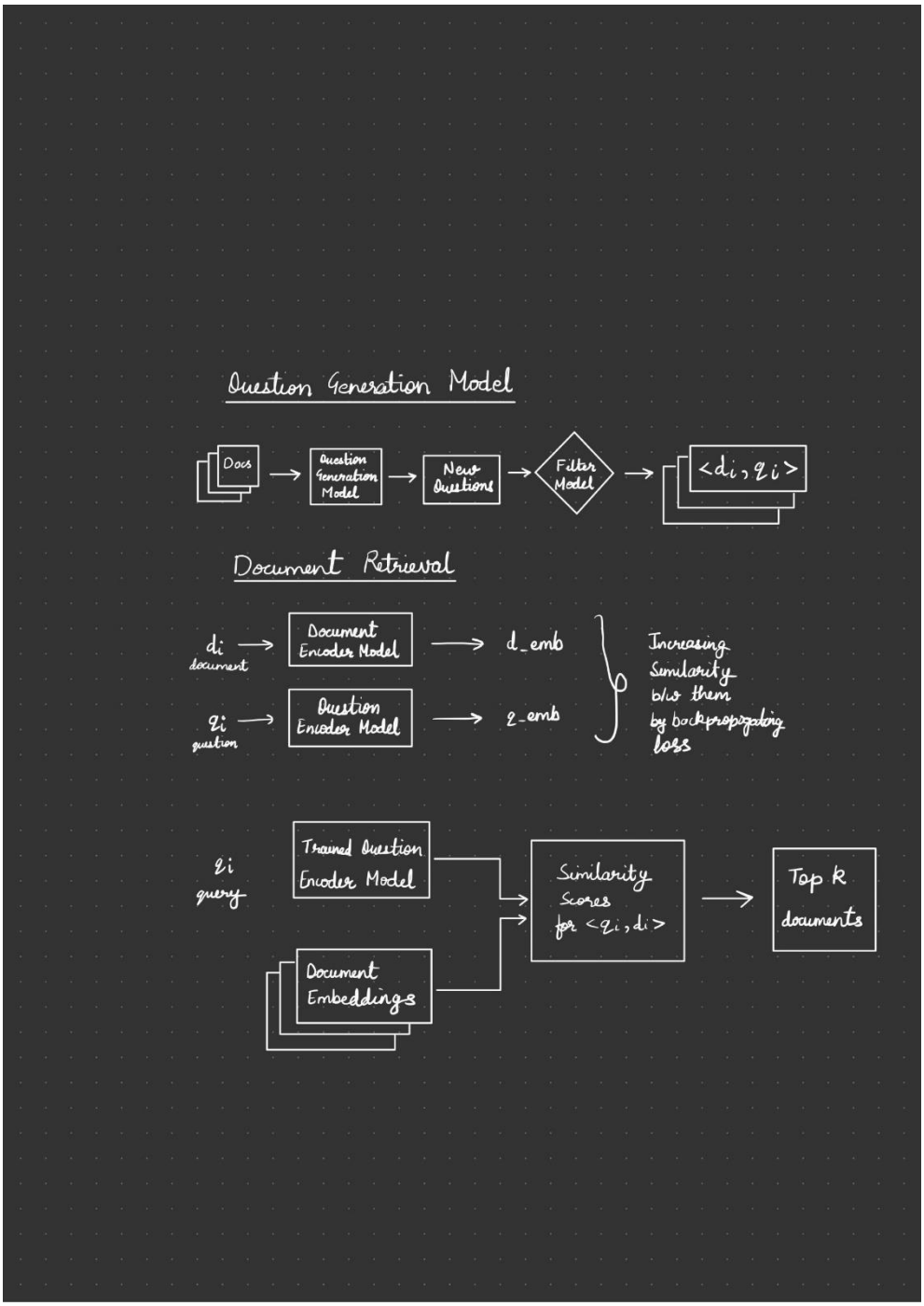
- **TRANSFER LEARNING APPROACHES:**
    - The author highlights the significance of feature vectors as numerical representations of words or sentences in neural network-based NLP models. **Pre-trained embeddings**, either word-level or subword-based,

have been instrumental in improving model performance in various NLP tasks. Subword embeddings address out-of-vocabulary issues and have proven effective for low-resource sequence labelling tasks. Recent trends involve pre-training large models like BERT and RoBERTa using language model objectives, benefiting low-resource languages with limited labelled data but abundant unlabeled data.

- Challenges persist, particularly in low-resource scenarios, where hardware requirements for large-scale models can be impractical. Studies suggest that smaller transformer sizes and alternative training strategies can yield comparable performance with fewer parameters. Additionally, data quality for low-resource languages, even in unlabeled datasets, may not match that of high-resource languages, impacting the effectiveness of pre-trained embeddings.
- The author underscores how language models, primarily trained on **general-domain data**, face challenges when applied to specialized domains with distinct language characteristics, causing a "domain gap." Fine-tuning models for target domains with additional domain-adaptive and task-adaptive pretraining using unlabeled data have shown performance gains in both high- and low-resource settings. Specialized domain-adapted language models, such as BioBERT and SciBERT for biomedical and scientific texts, demonstrate the effectiveness of this approach. Combining high-resource embeddings from the general domain with low-resource embeddings from the target domain and using attention-based meta-embeddings can further enhance domain-specific representations. Additionally, adversarial alignment of embeddings trained on diverse domains can generate domain-invariant representations.
- The author discusses how low-resource languages can benefit from labelled resources available in high-resource languages using **multilingual language representations**. These representations are achieved by training models on unlabeled corpora from various languages, like multilingual BERT and XLM-RoBERTa. In cross-lingual zero-shot learning, these models are applied to low-resource languages, leveraging task-specific labelled data from high-resource languages, notably for tasks like named entity recognition and reading comprehension. However, a substantial gap between low and high-resource settings persists. To bridge this gap, minimal amounts of target-task and language data have been proposed, significantly

boosting performance in low-resource languages. The alignment of languages in a common multilingual embedding space is also explored to improve transfer between languages. Despite these advancements, certain low-resource languages remain underrepresented in these models, particularly African and American languages, posing challenges for transfer learning.

PROPOSED ARCHITECTURE DESIGN:



4. Implementation Plan:

Task No.	Task	Estimated Completion Date
1	Literature Review	07/09/2023
2	Dataset Finalization	13/09/2023 - 18/09/2023
3	Question Generation Model	12/10/2023 - 18/10/2023
4	Training Document Retrieval Model	10/11/2023 - 15/11/2023
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.*

5. Interim and Final deliverables:

- Interim deliverables include progress reports and a codebase with partial implementations; mostly, we will be completing the **Question Generation Model**.

- 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.

References:

- 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