QS STARS RATED FOR EXCELLENCE 2012

FPT Fpt University

TRƯỜNG ĐẠI HỌC FPT

DETECTING DUPLICATE MULTIPLE CHOICE QUESTIONS (MCQS) IN THE LARGE QUESTION BANK

Team: AIP490-G8

Supervisor: Phan Duy Hung

Our Team

Phan Duy Hung

Advisor

Phan Duc Manh

Leader

Nguyen Tuan Minh

Member

Table of contents

01

Introduction

Motivation and Problem
Related Works
Contribution

02

Methodology

Pipeline Overview
OCR and SBERT
Similarity mesurement method

03

Experiment and Result

Model Implementation
Data collection
Result and analysis

04

Web Application Deployment

Propose adding method
Database
Role permissions

01.
INTRODUCTION

Contents

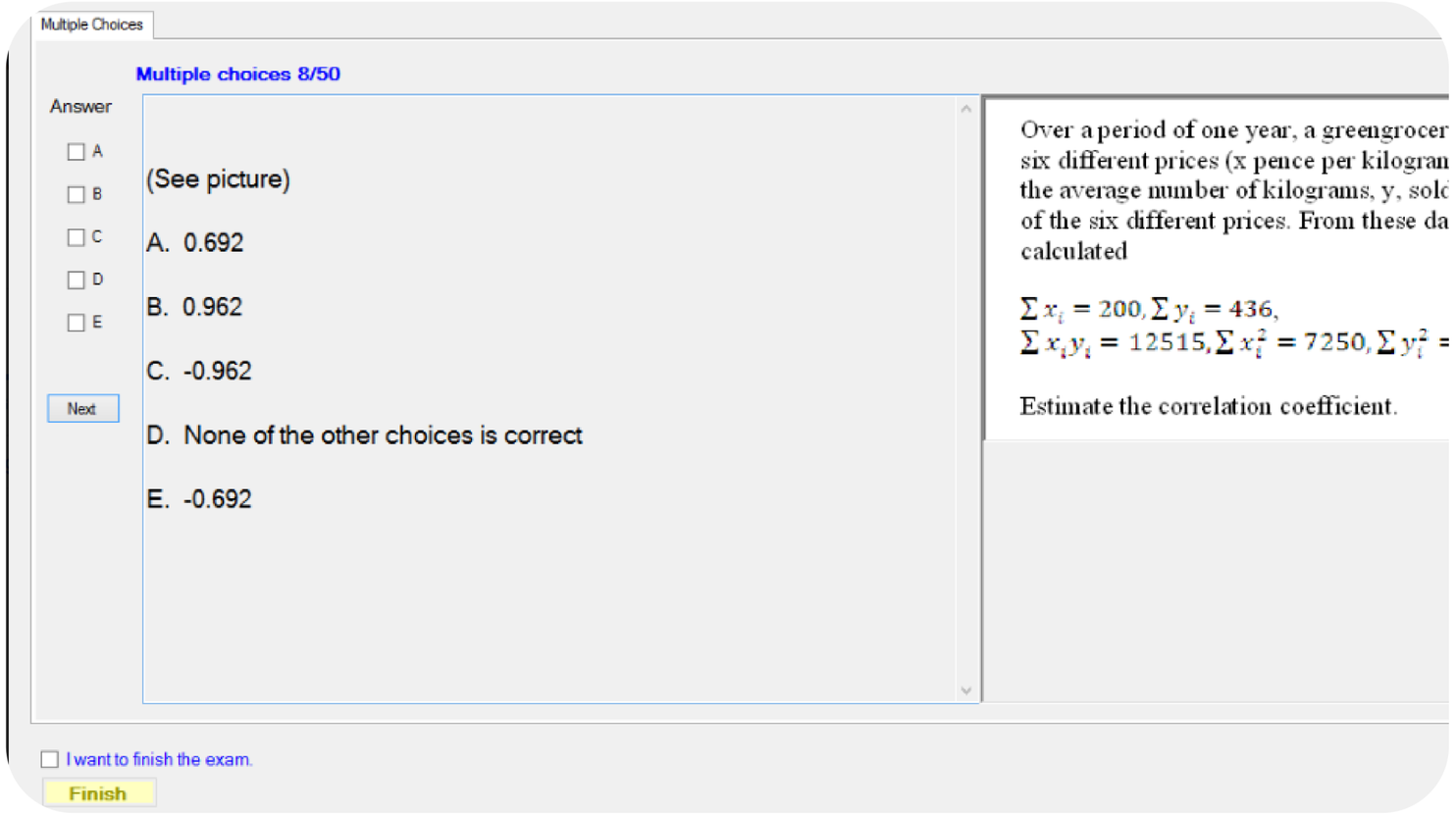
- Motivation and problem
- Related works
- Contribution

Motivation and Problem

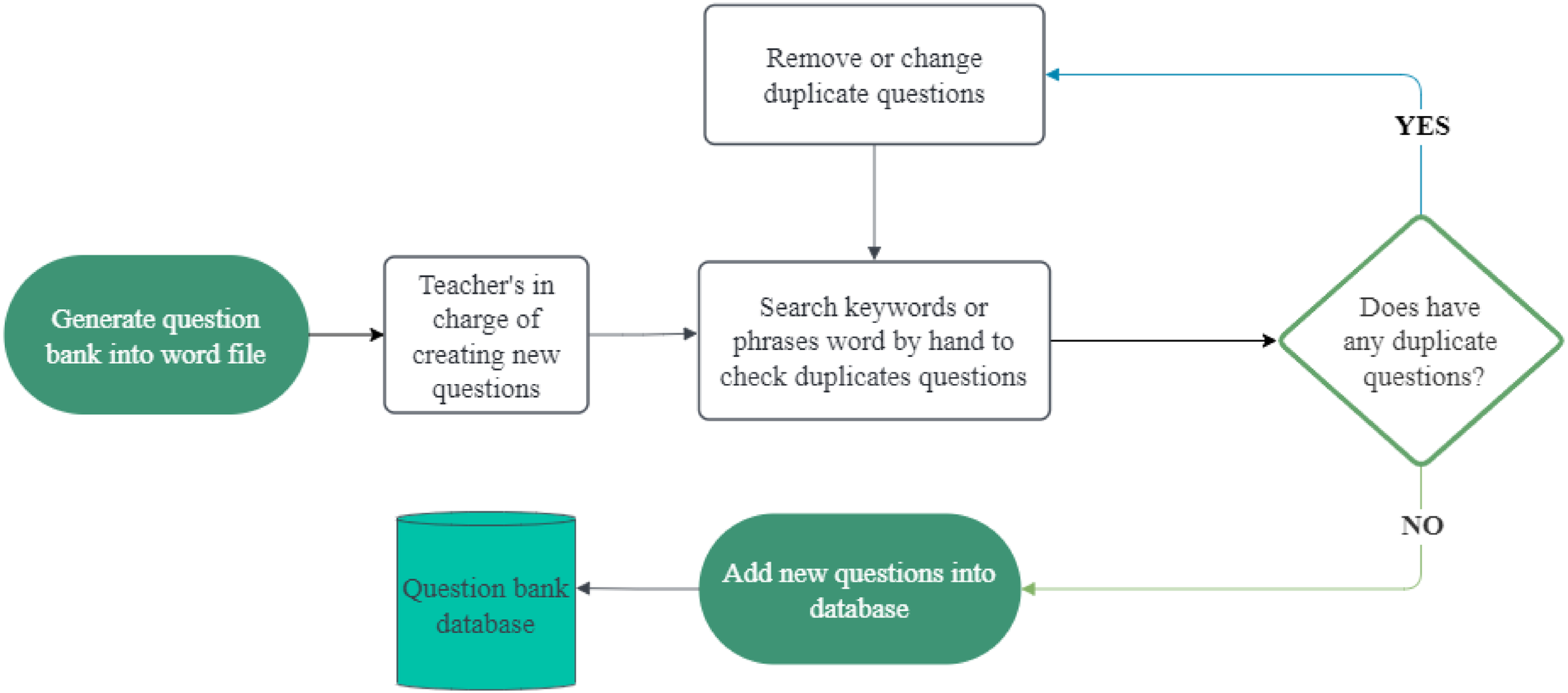
MCQs in FPTU

- The component of multiple-choice questions is:
  - **The question text**
  - **Answer options**
  - **Correct answers**

- Exist in two formats: text only and the combination of text and images.

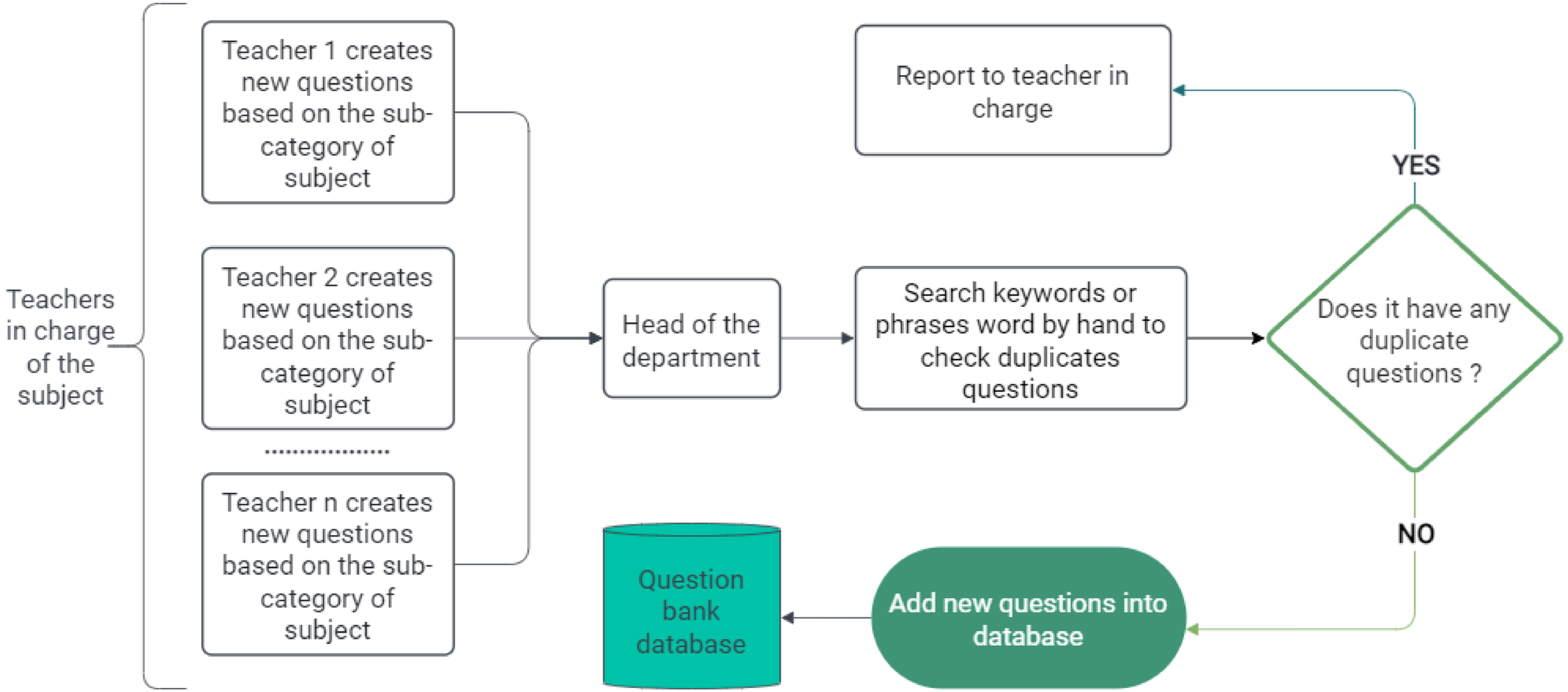


Motivation and Problem



Case one: Adding a new question to the question bank

Motivation and Problem



Case two: Adding a new question to the question bank

The problem of both adding cases

- In case one, the whole question bank must be imported again and every teacher can export the question bank to a Word file.
- When applying case two, it is hard to check duplicate questions on the variety of topics in the subject.
- The number of questions in database increase.
- Many questions have images, which makes the process of checking duplicate question become impossible.

Time-consuming when manually checking duplicate question and can not check duplicates in image format.

Related work

QUESTION BANK MANAGEMENT

A BRIEF TIMELINE

2012

Design of adaptive question bank development and management system

2018

Finding Similar Medical Questions from Question Answering Websites

2019

Question Bank Similarity Searching System (QB3S)

2020

Duplicate Question Detection With Deep Learning in Stack Overflow

Related work

TWO-SENTENCE COMPARISON

A BRIEF TIMELINE

2017

Sentence similarity detection in Malayalam language using cosine similarity

2018

Application of TF-IDF Feature for Categorizing Documents of Online Bangla Web Text Corpus

2020

Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks

Related work

OPTICAL CHARACTER RECOGNITION (OCR)

A BRIEF TIMELINE

2014

Reading Text in the Wild with Convolutional Neural Networks

2015

An End-to-End Trainable Neural Network for Image-Based

2016

TextBoxes: A Fast Text Detector with a Single Deep Neural Network

2017

EAST: An Efficient and Accurate Scene Text Detector

2019

Character Region Awareness for Text Detection

CONTRIBUTION

OUR MAIN TOPICS TODAY

IMPROVE QUESTION BANK MANAGEMENT

Developing a model that can identify similarities between questions, whether they are in text or image form

EVALUATE AND EXPERIMENT

Collect a Mcqs dataset for research purposes

BUILD WEB APPLICATIONS

The website is used to manage the question bank, the user can import the Mcqs and the website will detect duplicate questions that already exist in database

02. Methodology

Contents

- SBERT
- Similarity measurement method
- Evaluation metric
- OCR
- Overview Pipeline

SBERT

BASED ON

A modification of the BERT network

Use encoder transformation

ARCHITECTURE

Applied CLS-token, (MEAN-strategy, and MAX-strategy.

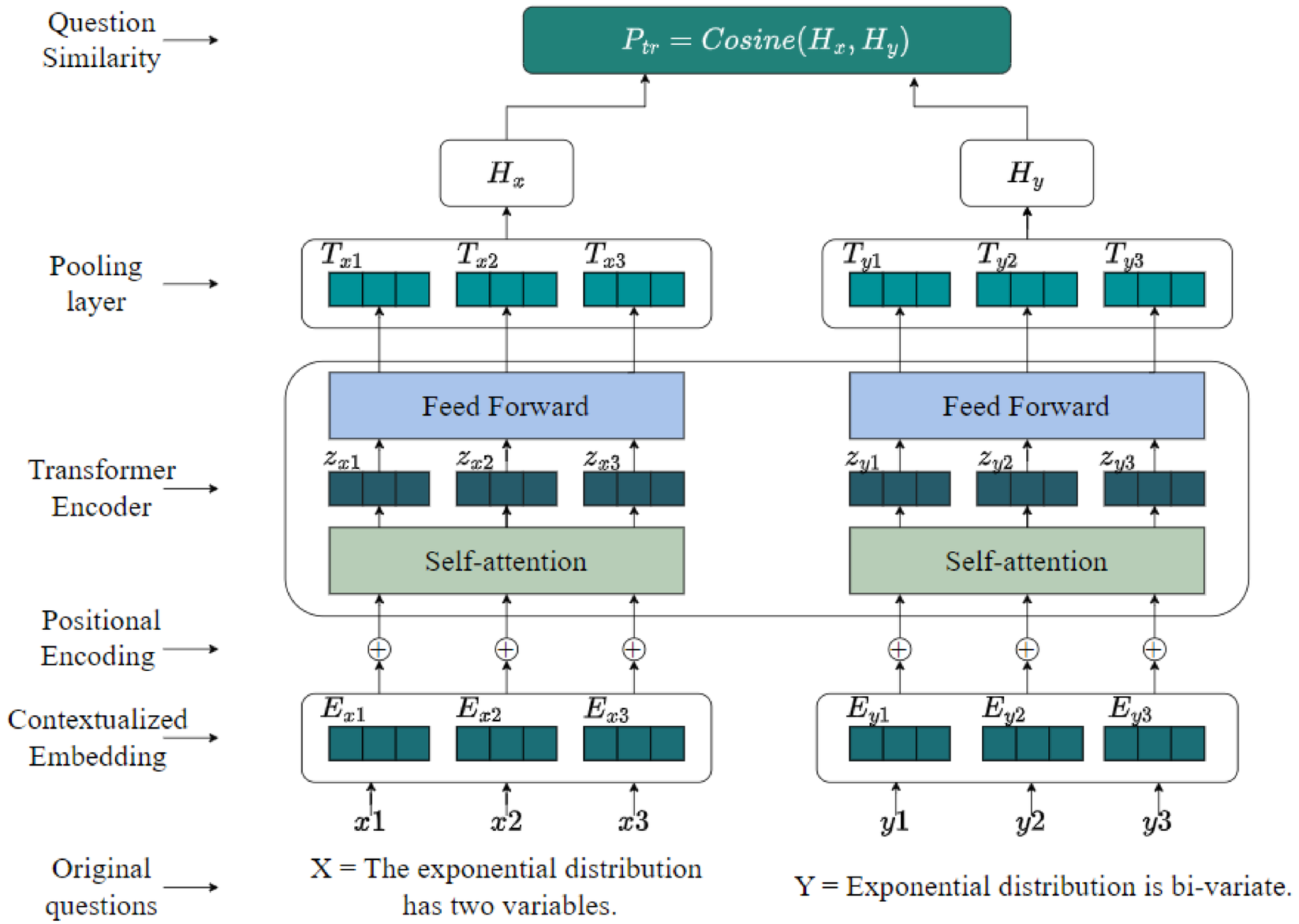
Use shared-weight siamese network structure to fine-tune

SOLVED ISSUES

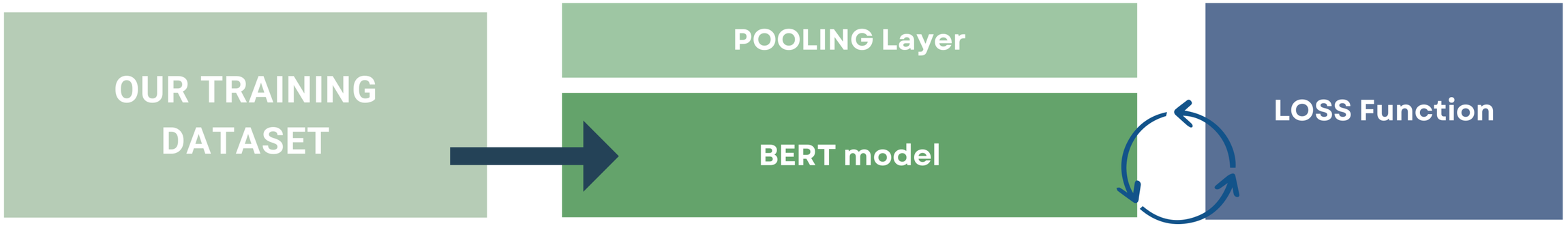
Efficient in terms of search time while preserving the accuracy of BERT on STS tasks

Has the ability to balance the context in both the left and right directions in a sentence base on the transformer technique.

SBERT model


Question Similarity
$P_{tr} = Cosine(H_x, H_y)$
H_x
H_y
Pooling layer
T_{x1} T_{x2} T_{x3}
T_{y1} T_{y2} T_{y3}
Transformer Encoder
Feed Forward
Feed Forward
z_{x1} z_{x2} z_{x3}
z_{y1} z_{y2} z_{y3}
Self-attention
Self-attention
Positional Encoding
Contextualized Embedding
E_{x1} E_{x2} E_{x3}
E_{y1} E_{y2} E_{y3}
$x1$ $x2$ $x3$
$y1$ $y2$ $y3$
Original questions
X = The exponential distribution has two variables.
Y = Exponential distribution is bi-variate.

Fine-tune Sbert



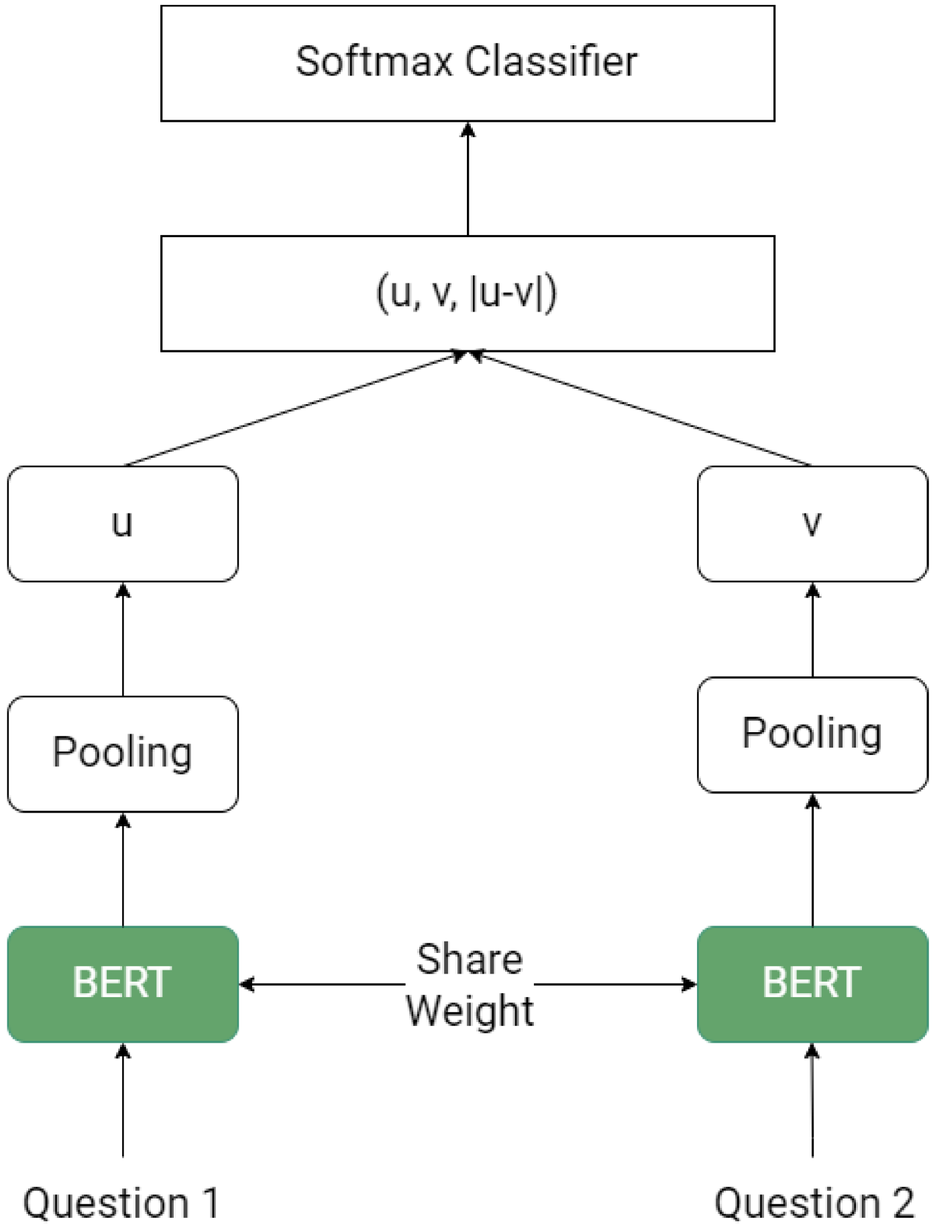
- Applied shared-weight siamese network structure after BERT follow by pooling layer and using three losses softmax, triplet loss, multiple negative loss to fine-tune
- SNN focuses on learning embeddings in the deeper layer that place the same classes close together so that it suitable for the Semantic Textual Similarity task.

SBERT with cross-entropy loss function

- Applied by adding a softmax layer and a cross-entropy loss to learn weight and directly predict a label

- Concatenate two sentence embeddings and multiply by trainble weight:

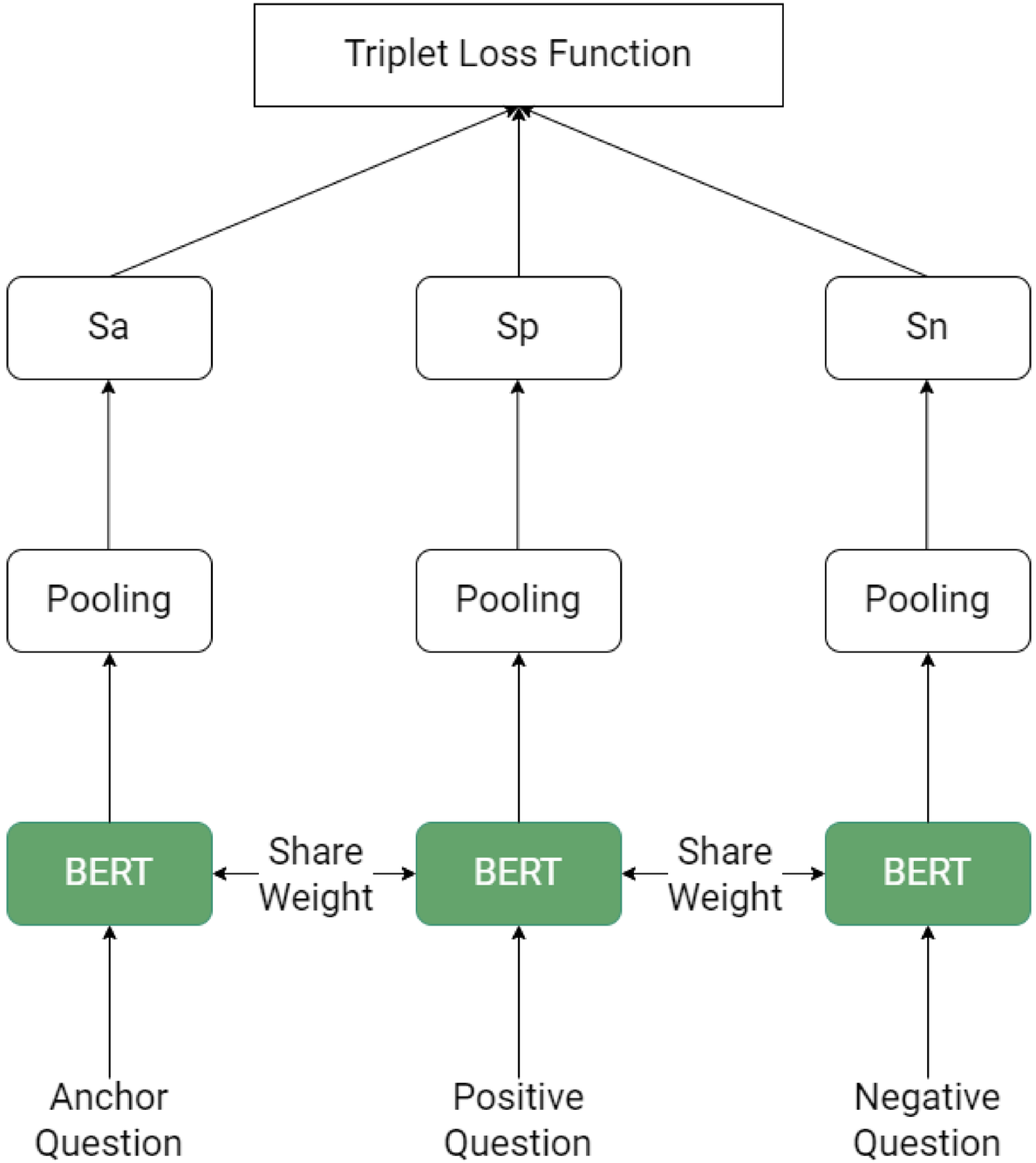
$$L = softmax(W_t, (u, v, |u - v|))$$



SBERT with triplet loss function

- Employing the anchor question, the positive question, and the negative question
- Reducing the distance of the anchor question with positive question embedding and increase the distance with negative question embedding simultaneously.

$$L = \sum_{i=1}^{N} max(\|s_a - s_p\| - \|s_a - s_n\| + m, 0)$$

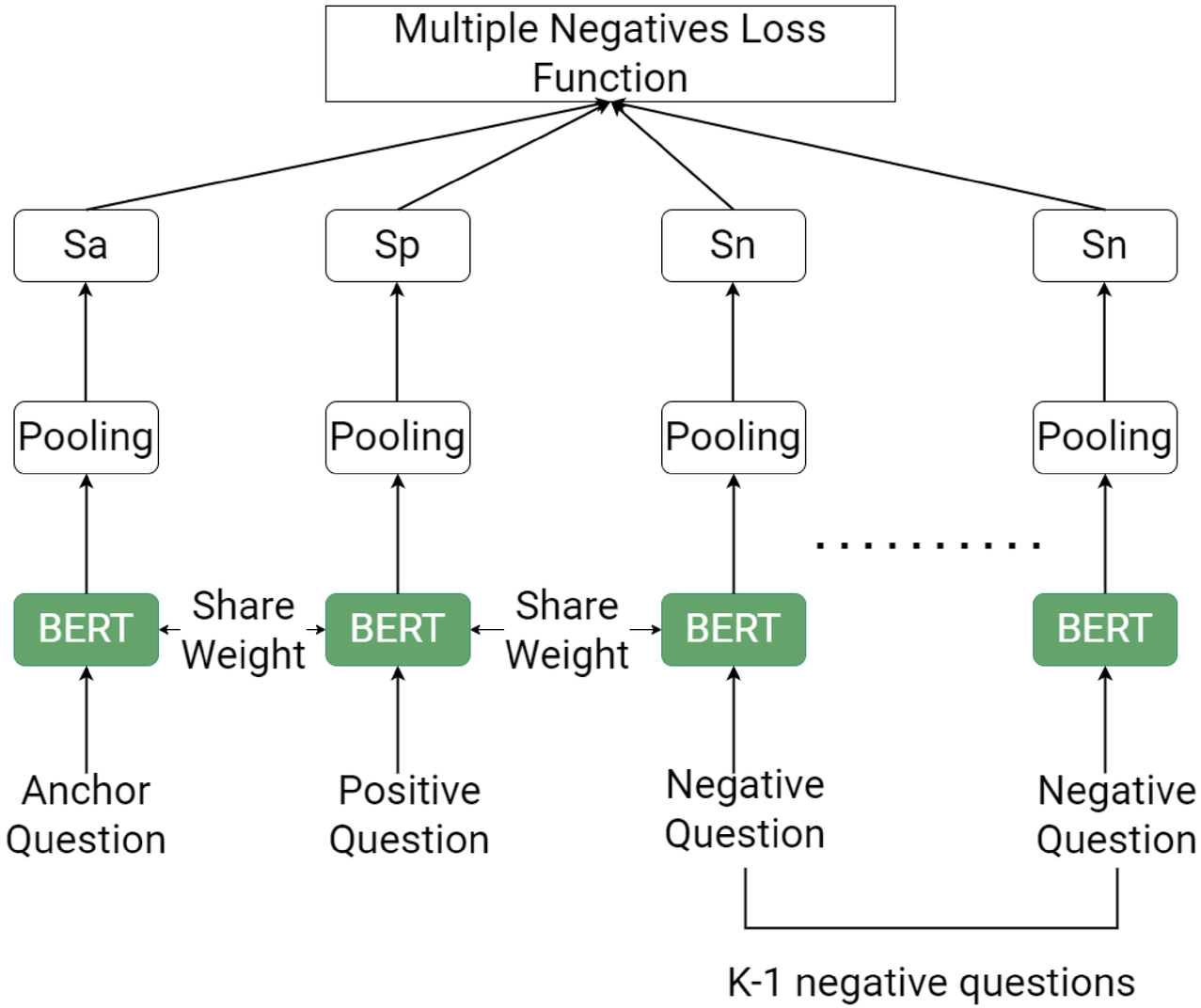


SBERT with multiple negatives loss function

- Focuses on balancing the distance of a positive question over multiple negative questions simultaneously

- Suitable for skewed data

$$L = -\frac{1}{K}\sum_{i=1}^{K}[S(x_i, y_i) - log\sum_{j=1}^{K} e^{S(x_i, y_j)}]$$



Similarity measurement methods

01 **COSINE**

$$cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

02 **EUCLID**

$$d(\overrightarrow{A_i}, \overrightarrow{B_i}) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

03 **SPEARMAN**

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} (A_i - B_i)^2}{n(n-1)^2}$$

Evaluation method

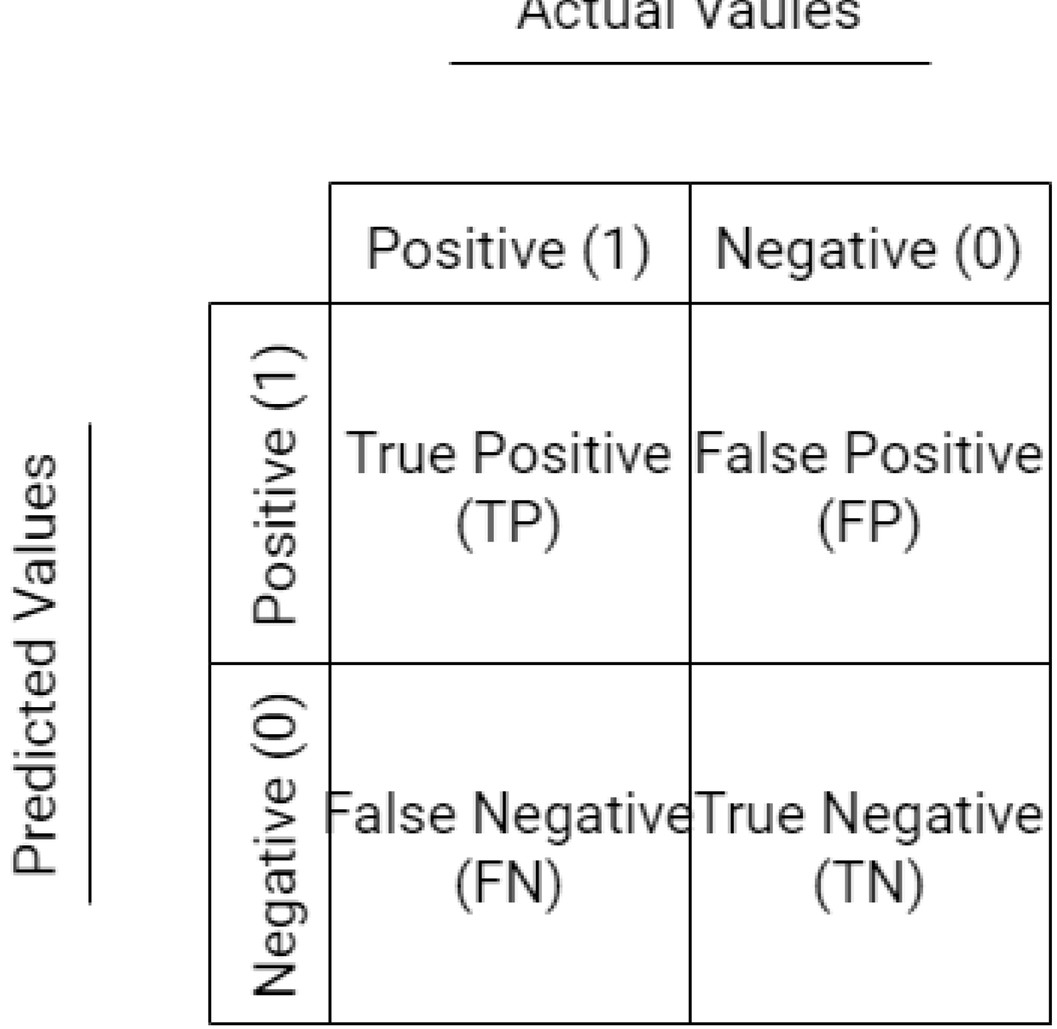
F1 SCORE

focuses on a model's performance inside each class

$$Precision = \frac{TP}{TP + FP}$$

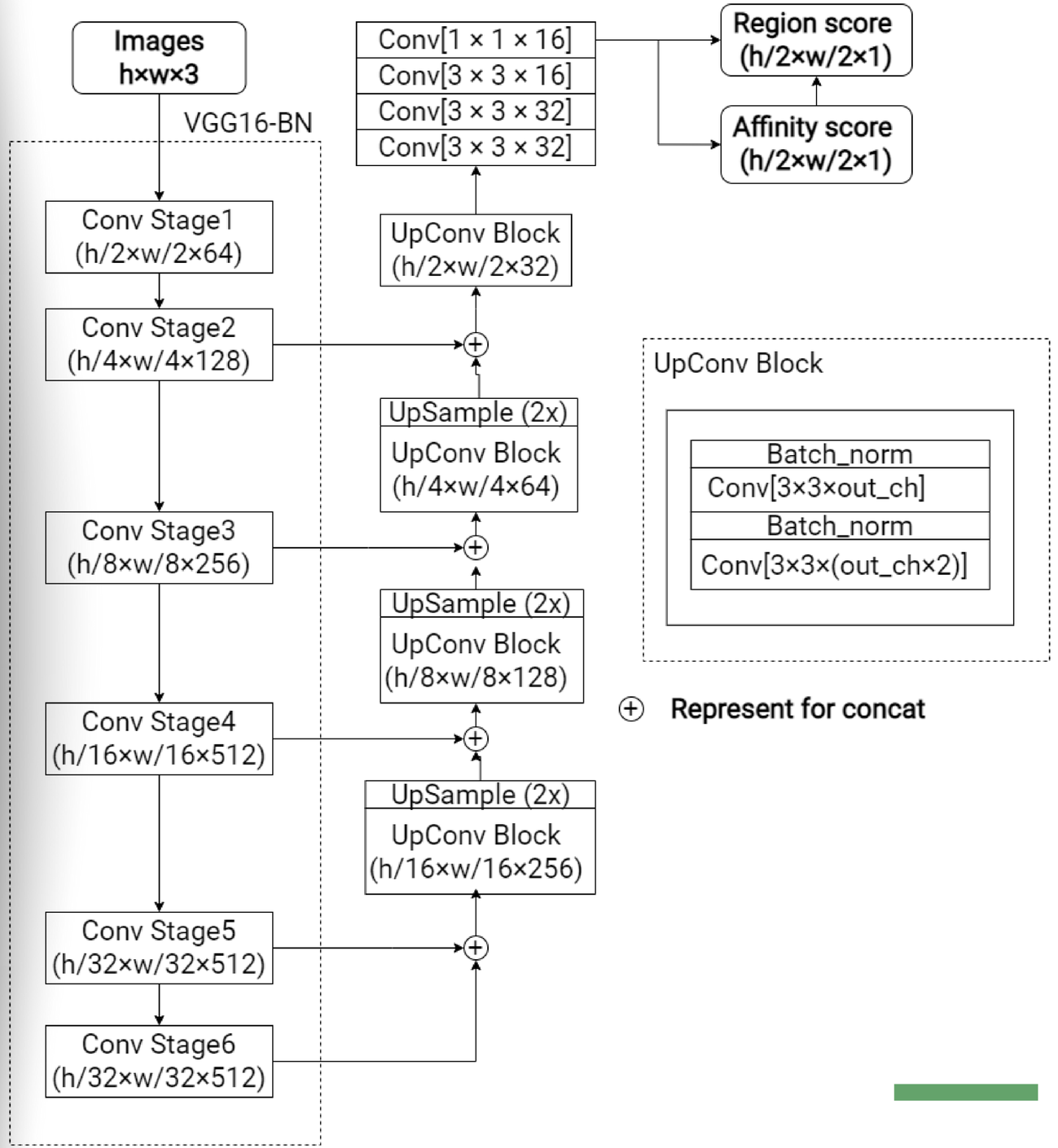
$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precison \times Recall}{Precision + Recall}$$



EasyOCR

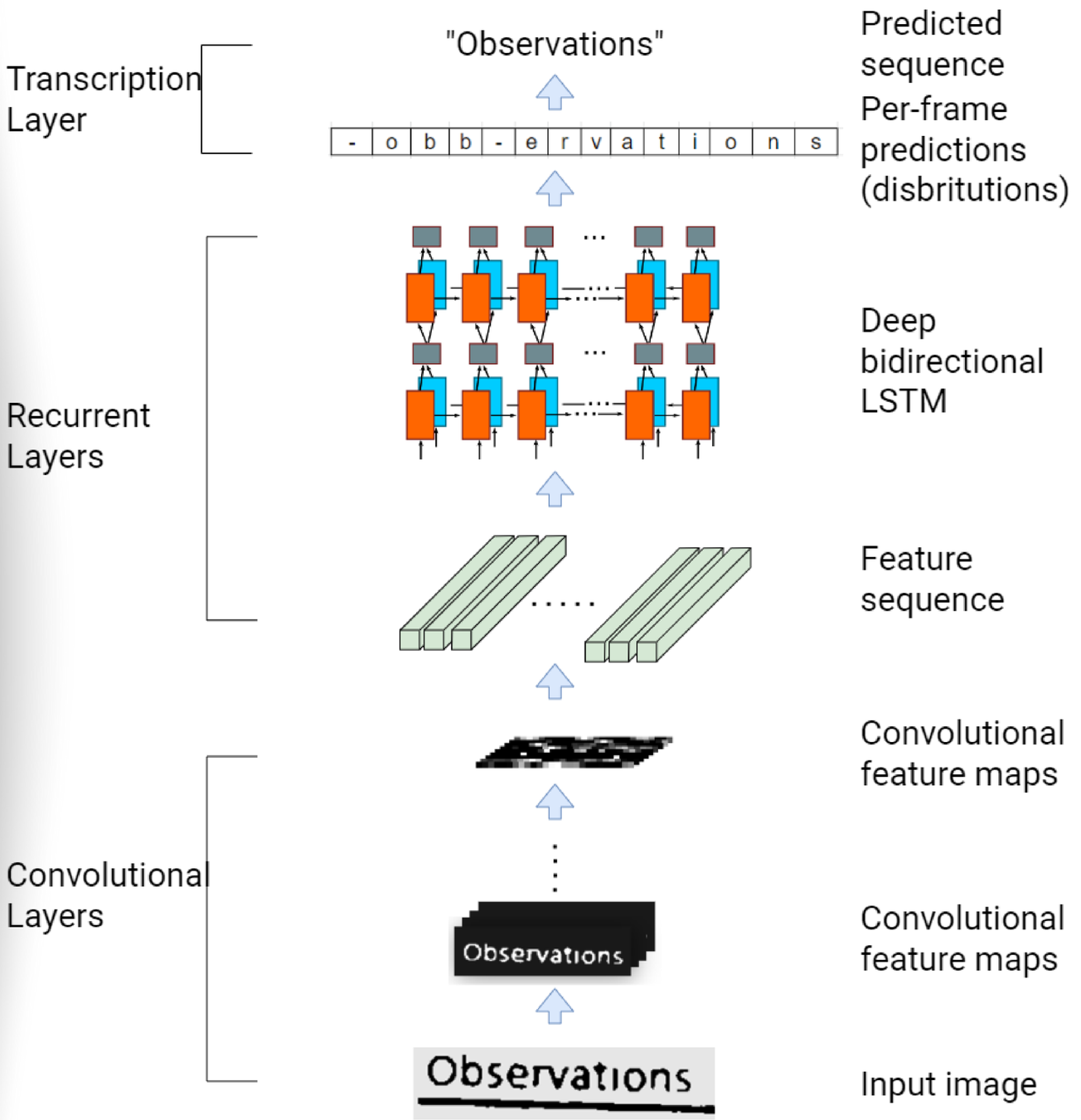
- Use CRAFT to localize the character region in the image and group detected characters to a text instance
- Trained in a fully convolutional network architecture based on VGG-16 and returns the region score and affinity score
- The region score and affinity score are used to localize each individual character in the image are the out put of next stage.



EasyOCR

- the input for CRNN is the bounding boxes and the objective is to recognize the letter inside the bounding boxes

- CRNN is made up of three parts: convolutional layers, recurrent layers, and a transcription layer



EasyOCR

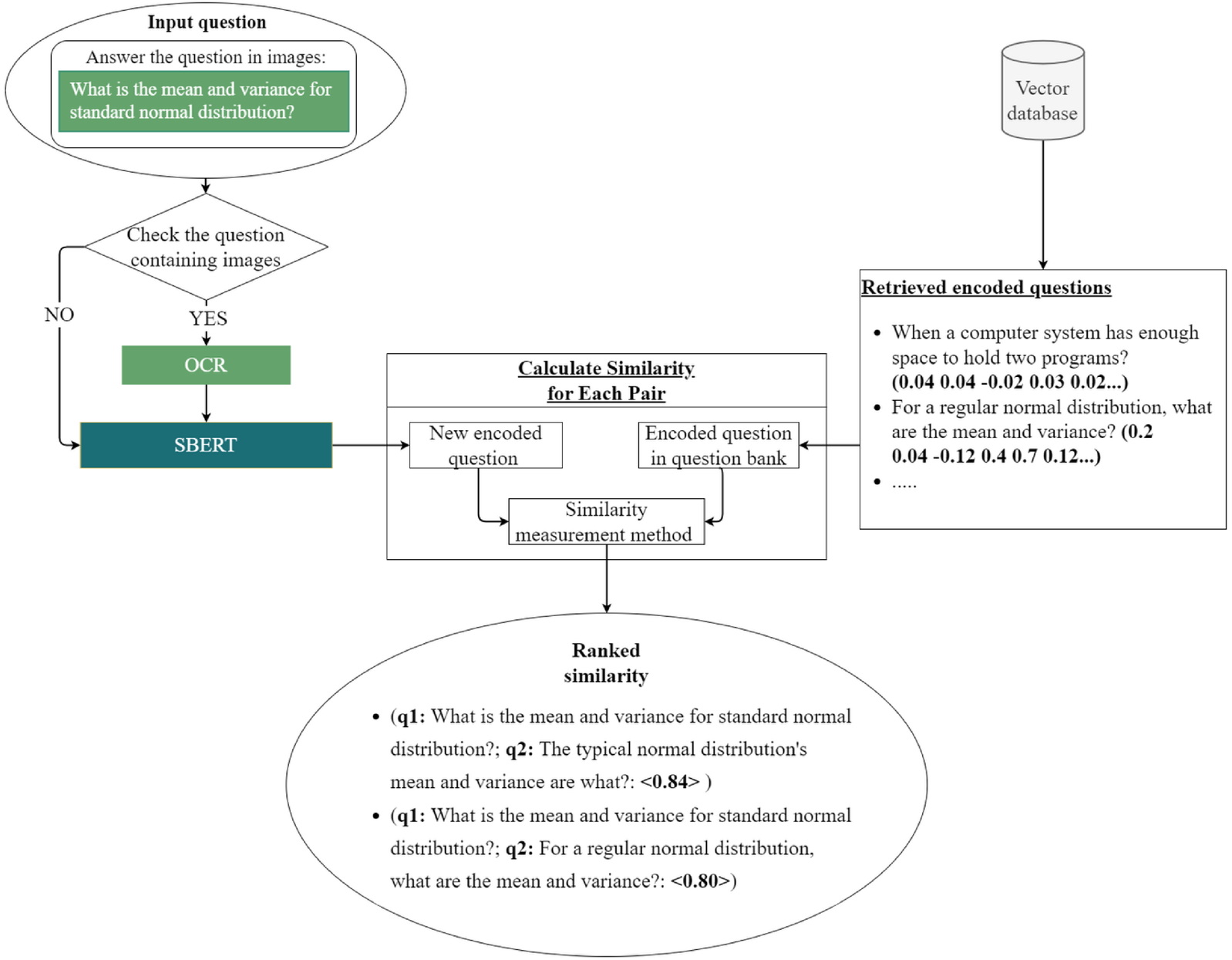
Which of the given options provides the increasing order of asymptotic complexity of functions f1, f2, f3 and f4?

f1(n) = 2^n
f2(n) = n^(3/2)
f3(n) = nLogn
f4(n) = n^(Logn)

a) f3, f2, f4, f1
b) f3, f2, f1, f4
c) f2, f3, f1, f4
d) f2, f3, f4, f1

The output of EasyOCR: Which of the given options provides the increasing order of asymptotic complexity of functions fl, f2, f3 and f4?
f1(n) = 2^n f2(n) = n^(3/2) f3(n) = nLogn f4(n) = n^(Logn)
a) f3, f2, f4, f1b) f3, f2, f1, f4 c) f2, f3, f1, f4 d) f2, f3, f4, f1

Overview Deployment



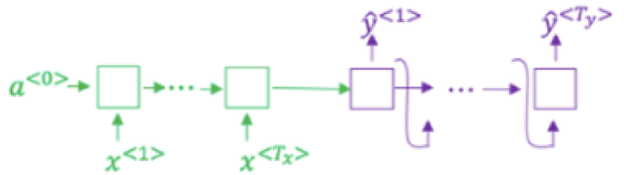
03.
Experiment and Result

Contents

- Data collection
- Implementation
- Result and analysis

Data collection

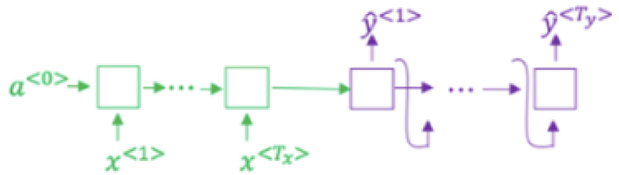
- We collected data from 1242 multiple-choice questions based on the FPT university question bank,
- The datasets include two types: Math and English
- The number of English questions are 682 and Math are 560

QN= im234	Let z have the following value Consider using this encoder-decoder model for machine translation.  This model is a "conditional language model" in the sense that the encoder portion (shown in green) is modeling the probability of the input sentence x. [file: im234.png]
a.	homoscedasticity
b.	Normality of errors
c.	Independence of errors
d.	Linearity of relationship
answer	D

QN= eng20	When seeking to resolve group tensions, the dialectic approach recommends a _____ approach.
a.	compromise
b.	either/or
c.	both/and
d.	approach/avoidance
e.	task/maintenance

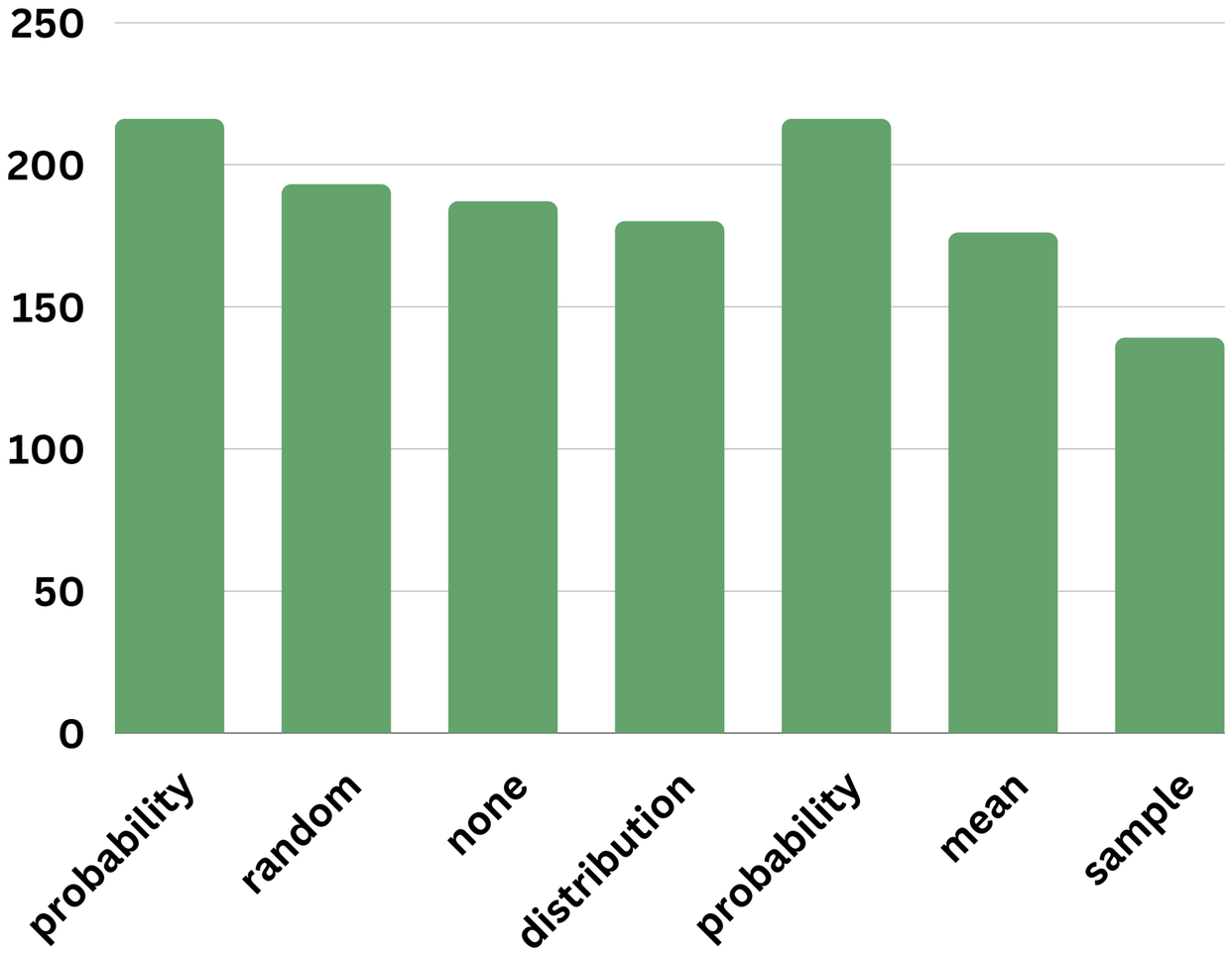
Data statistic

Information	Value
Corpus size	312811
Vocabulary size	10306
Average length of question	252

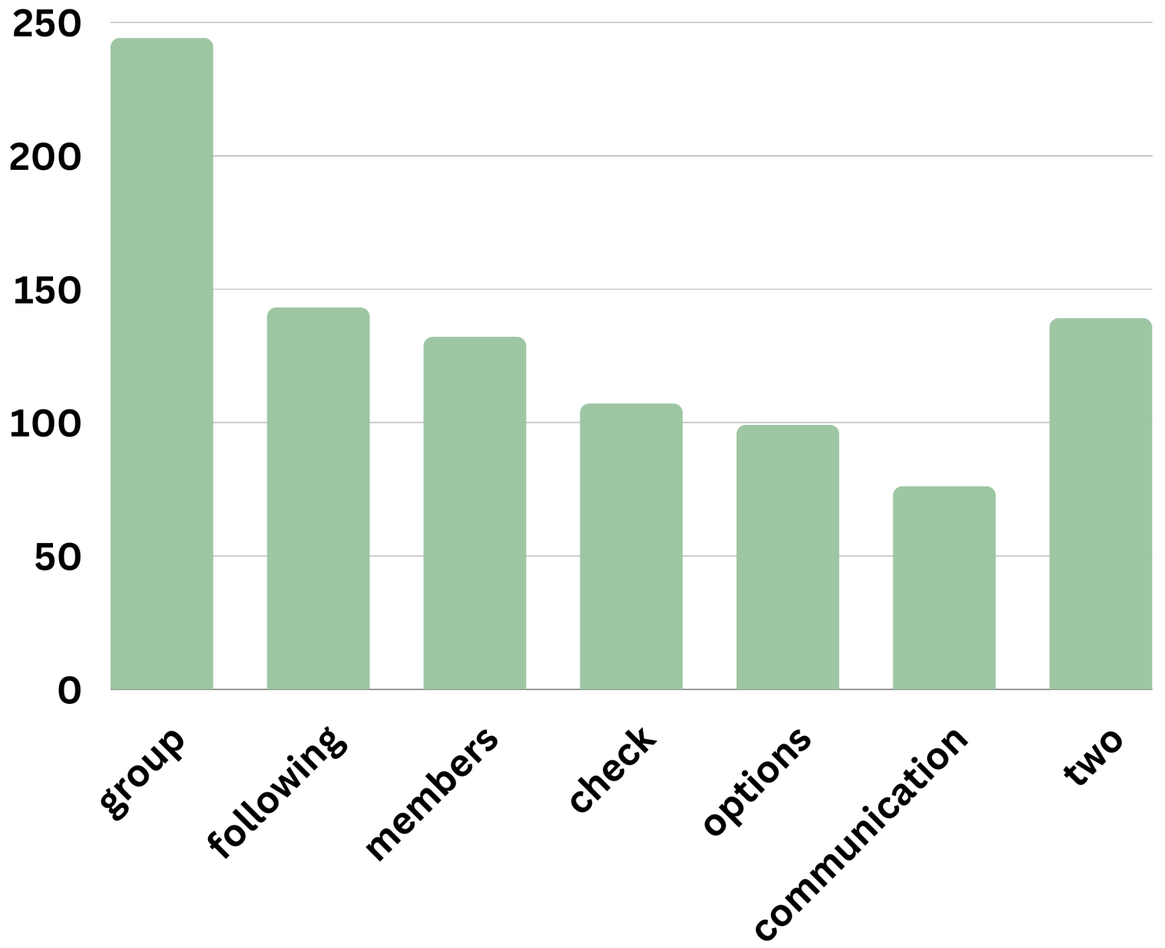
QN= im234	Let z have the following value Consider using this encoder-decoder model for machine translation.  This model is a "conditional language model" in the sense that the encoder portion (shown in green) is modeling the probability of the input sentence x. [file: im234.png]
a.	homoscedasticity
b.	Normality of errors
c.	Independence of errors
d.	Linearity of relationship
answer	D

QN= eng20	When seeking to resolve group tensions, the dialectic approach recommends a _____ approach.
a.	compromise
b.	either/or
c.	both/and
d.	approach/avoidance
e.	task/maintenance

Data collection



Word frequency for Math questions types



Word frequency for English questions types

Implementation

1

FRAMEWORK

Machine learning is implemented in PyTorch

2

DEVICE

The experiments were conducted on a GPU Nvidia Tesla T4 15Gb GPU

3

FINE-TUNE MODEL

- Apply two pre-trained models SBERT
- Train models on our datasets with 5 epochs, and batch size of 8

ALL-MPNET-BASE-V2

- Trained on a 1 billion sentence pairs dataset
- Sequence length limited to 128 tokens, batch size of 1024, and learning rate of 2e-5 with 100,000 steps

MULTI-QA-MPNET-BASE-DOT-V1

- The concatenation from multiple datasets to fine-tune model. In total Tranning data have about 215M (question, answer) pairs

Result and Analysis

Model	Cosine	Spearman	Euclid
all-mpnet-base-v2	0.89	0.9	0.9
multi-qa-mpnet-base-dot-v1	0.89	0.9	0.12
all-mpnet-base-v2 fine-tune softmax loss	0.75	0.76	0.76
multi-qa-mpnet-base-dot-v1 sofmax loss	0.77	0.77	0.44
all-mpnet-base-v2 fine-tune triplet loss	0.80	0.80	0.80
multi-qa-mpnet-base-dot-v1 triplet loss	0.88	0.89	0.25
all-mpnet-base-v2 multiple negative loss	**0.94**	**0.95**	**0.93**
multi-qa-mpnet-base-dot-v1 multiple negative loss	0.94	0.92	0.15

Accepatation letter of conference

Acceptance Notification : Intelligent Systems Conference (IntelliSys) 2023

Inbox ×

IntelliSys Conference <IntelliSys@saiconference.com>
to hungpd2, manhpdhe150883, me

18:48 (1 hour ago)

Dear Phan Duy Hung, Phan Duc Manh, Nguyen Tuan Minh,

Congratulations, your submitted paper "Detecting Duplicate Multiple Choice Questions in The Large Question Bank" has been reviewed and accepted for presentation at the Intelligent Systems Conference (IntelliSys) 2023, to be held on 7 & 8 September 2023 in Amsterdam, The Netherlands (Van der Valk Hotel Amsterdam - Amstel).

Intelligent Systems Conference (IntelliSys) 2023 will focus on areas of intelligent systems and artificial intelligence and how it applies to the real world. IntelliSys provides a leading international forum that brings together researchers and practitioners from diverse fields with the purpose of exploring the fundamental roles, interactions as well as practical impacts of Artificial Intelligence.

In order to attend, present and publish your paper, please register online at https://saiconference.com/IntelliSys2023/Register (Registration closes March 31, 2023). We also accept fees via bank/wire transfer.

IntelliSys 2023 proceedings will be published in the Springer series "Lecture Notes in Networks and Systems" and submitted for consideration to Web of Science, SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH and SCImago.

The Conference Board has decided that the reviewers' feedback and invitation letter for visa applications (if necessary) will be emailed to the author(s) after the registration process. If you would like to receive the reviewer feedback for representation at your university/organization, please feel free to contact us.

Again, congratulations and I look forward to your participation!

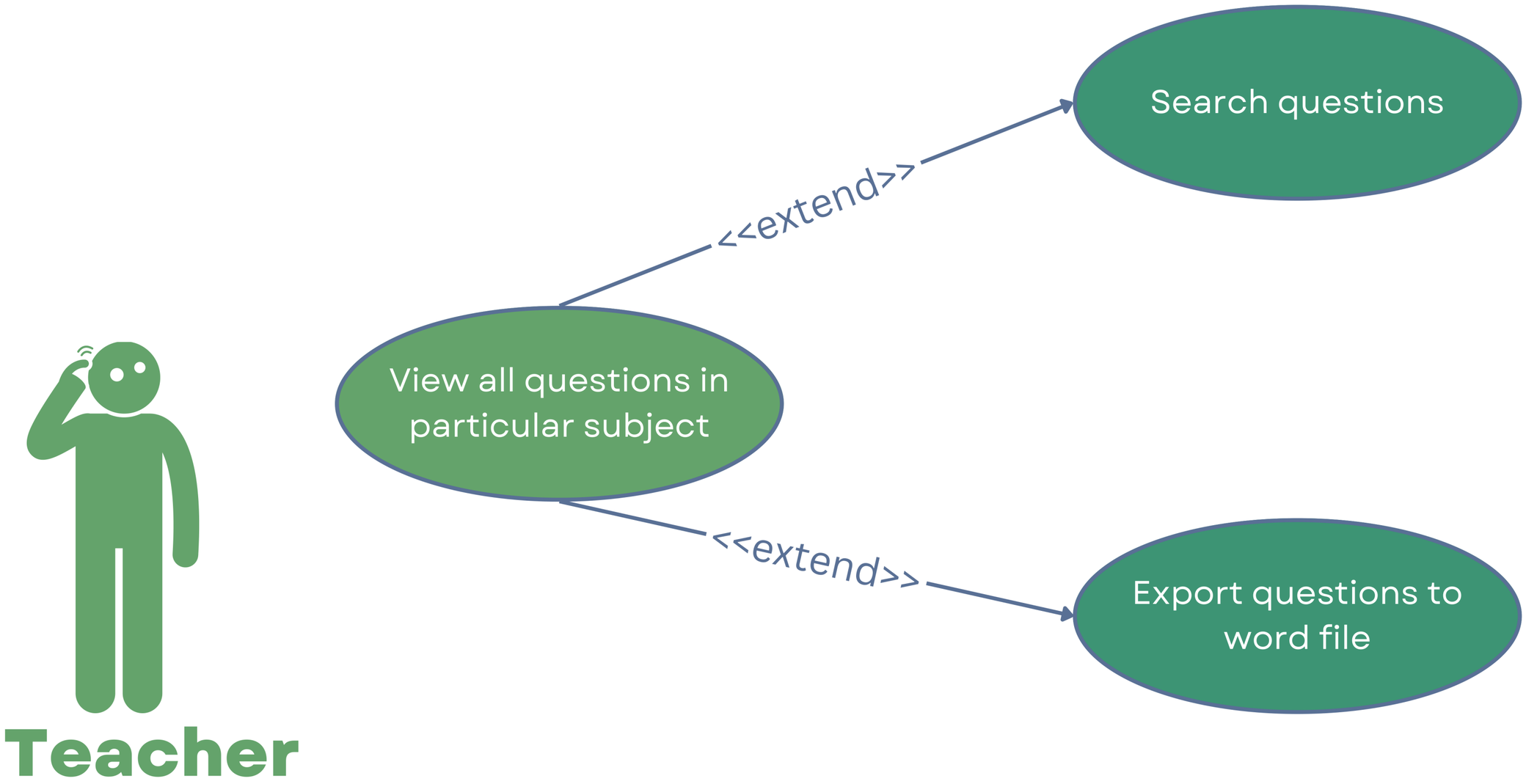
Regards,
Conference Manager
IntelliSys Conference

04.
Web Application Deployment

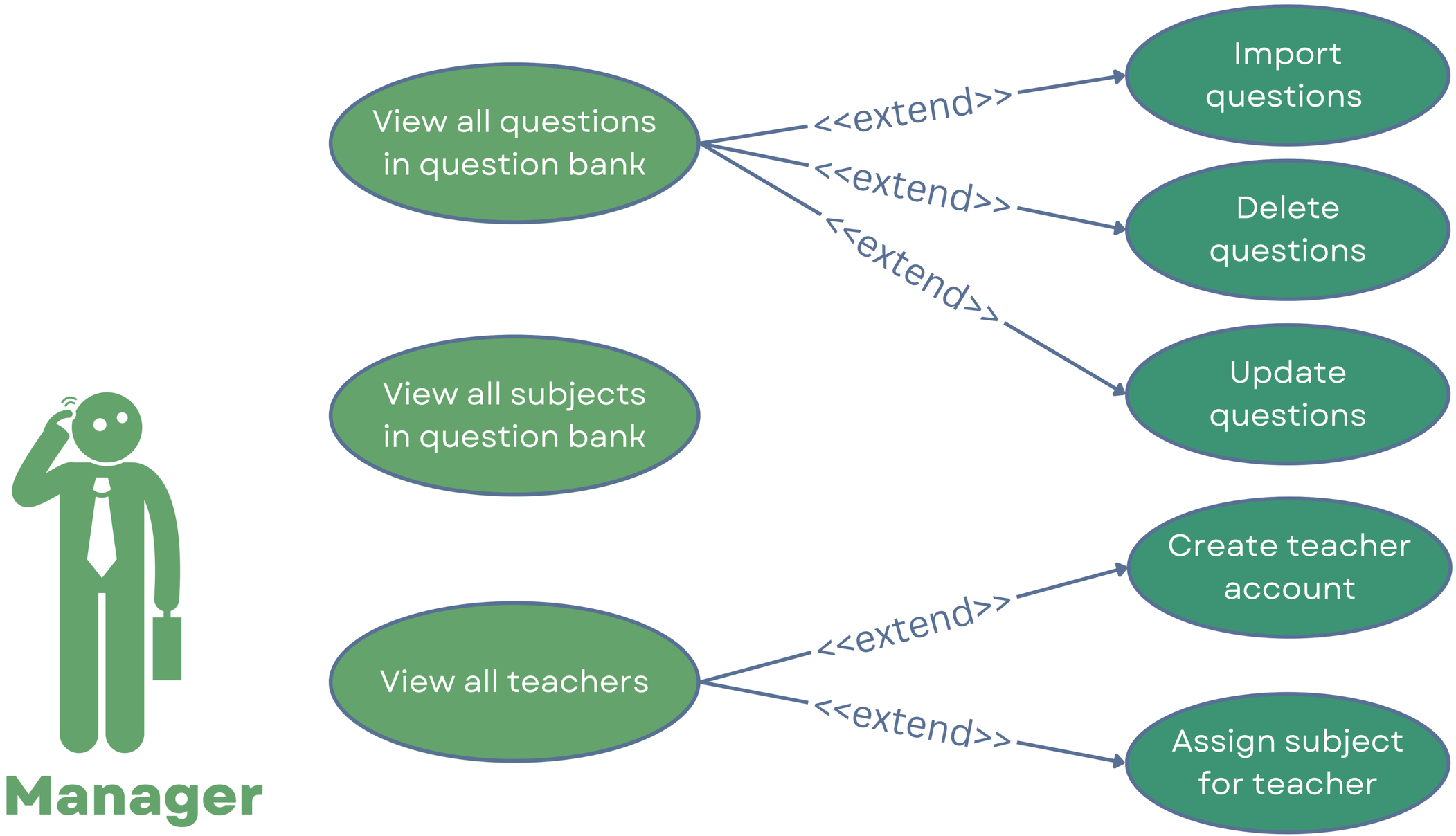
Contents

- Usecase Specifications
- Database ERD
- Result and analysis

Use case specifications



Use case specifications

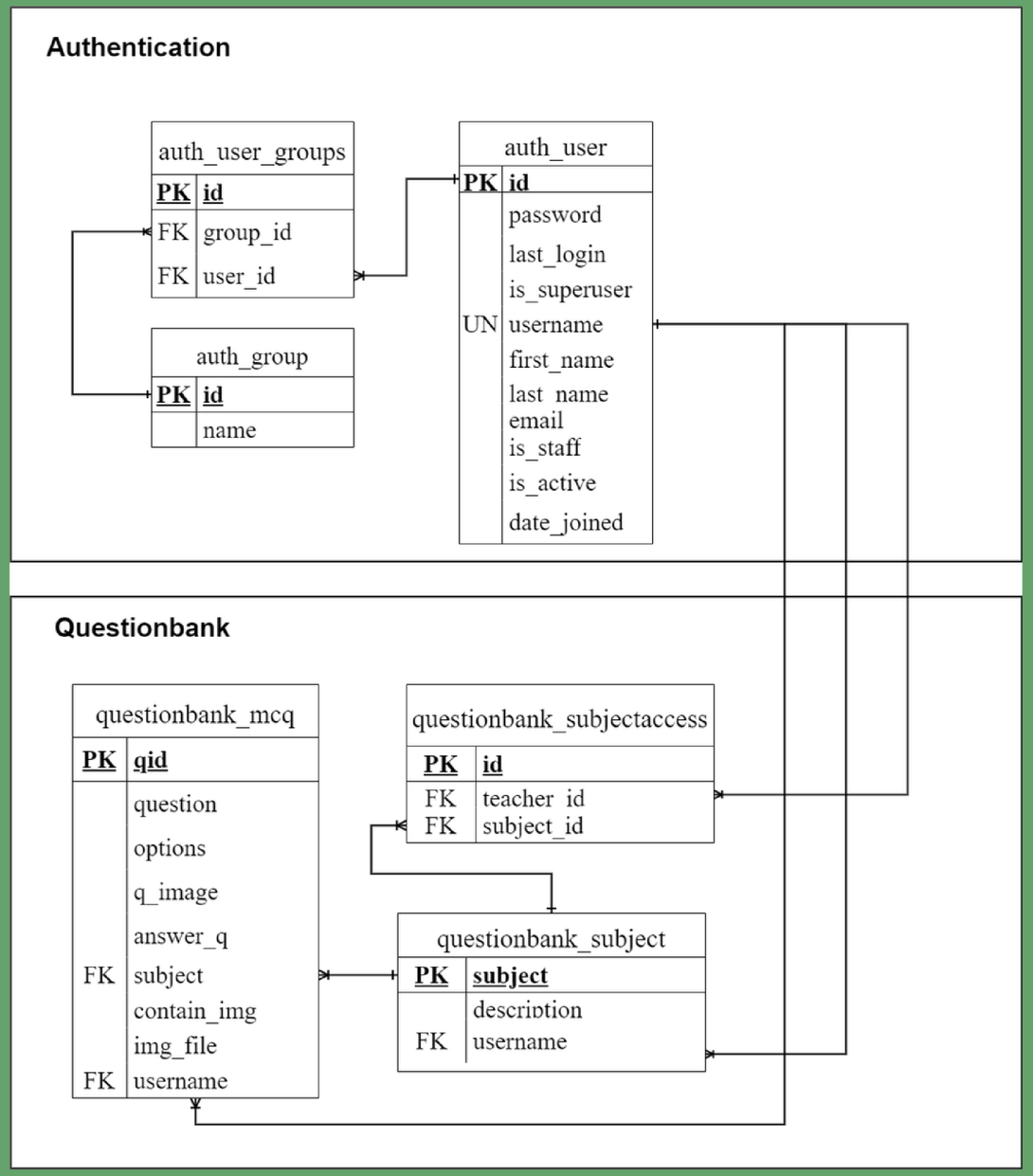


Database ERD

- The tables auth_group, and auth_user_group divide users into two main groups with different roles including teacher and manager

- .The table auth_user contains data of all users with the unique filed username for login to the system

- Each user must be assigned to access the subject by the manager and store it in the questionbank_subjectaccess table to interact with specific subject in the question bank



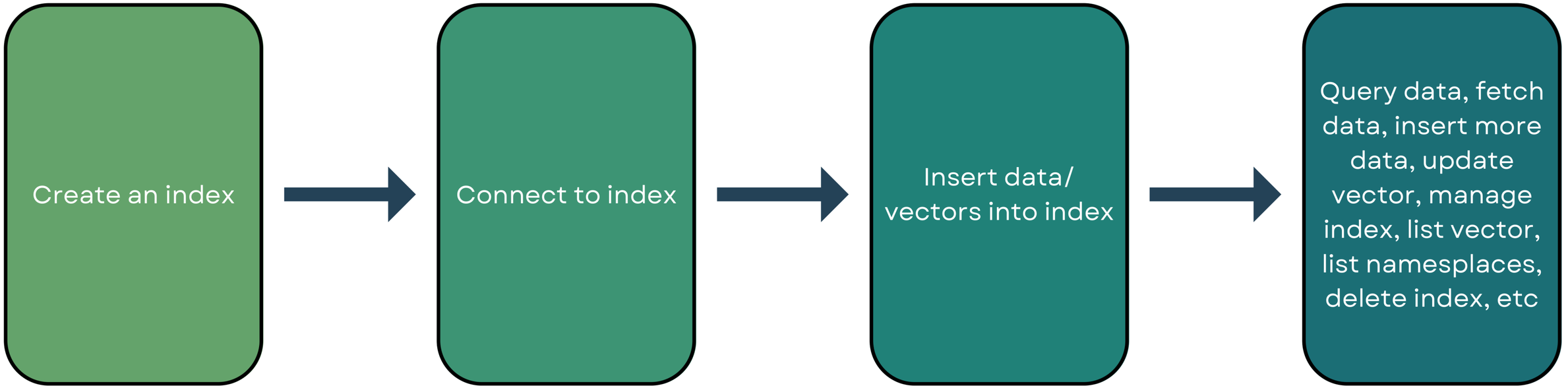
Pinecone

EMBEDDING DATABASE

LONG-TERM MEMORY FOR AI

- Pinecone 30 is a vector database, which has features like CRUD operations, metadata filtering, horizontal scaling, indexes, and saves vector embeddings for fast retrieval and similarity searches

- The Pinecone vector database makes it easy to build high-performance vector search applications. Developer-friendly, fully managed, and easily scalable without infrastructure hassles.



CONCLUSIONS

This study proposed a method for assisting the administration of the question bank at FPT University focused on detecting question duplication with a model F1 score's 0.95. In future works, we aim to tackle the problem of question duplication on an actual question bank of FPT University with different languages such as Vietnamese, Chinese, and Japanese. The presence of audio-based multiple-choice questions is also taken into consideration.

Thank you!