AI TRAINING PROGRAM - VINBIGDATA

NATURAL LANGUAGE PROCESSING
PROJECT REPORT

VIETNAMESE LEGAL TEXT RETRIEVAL

GROUP 8

Lê Trần Thắng - 3733351
Bùi Mạnh Cường - 3733351
Trần Khánh Lương - 3733351
Nguyễn Nho Trung - 3733351
Nguyễn Nhật Quang - 3733351

Supervisors
Assoc. Prof. Lê Thanh Hương
Assoc. Prof. Bùi Thị Mai Anh
Assoc. Prof. Nguyễn Kiêm Hiếu

Hanoi, 2023

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our supervisors Assoc. Prof. Lê Thanh Hương, Bùi Thị Mai Anh and Lê Kiêm Hiếu who carefully guided us during the period of working on this project. We know that there are so many mistakes and limitations in our work, but thanks to all the patience, carefulness and enthusiastic of you, we can overcome them and gain much crucial knowledge.

Contents

1 Introduction 3
1.1 Overview . . . 3
1.2 Problem formulation . . . 3
1.2.1 Legal documents corpus . . . 3
1.2.2 Question-Answer set . . . 4

2 Theoretical preliminaries 6
2.1 BM25+ . . . 6
2.1.1 BM25 . . . 6
2.1.2 BM25+ . . . 7
2.2 PhoBERT . . . 7
2.3 Dual Encoder and Cross Encoder in information retrieval . . . 8
2.3.1 Dual Encoder (Bi-Encoder): Dense Passage Retrieval[5] . . . 8
2.3.2 Cross-Encoder classifier . . . 9

3 The two-stage model for legal text retrieval problem 10
3.1 Two-stage retrieval system . . . 10
3.2 Processing technique . . . 11
3.2.1 PhoBERT processing . . . 11
3.2.2 BM25+ processing . . . 13
3.3 Bi-encoder retriever . . . 13
3.3.1 Number of hard negative documents . . . 13
3.3.2 Finetuning before training . . . 14
3.3.3 Training techniques . . . 14
3.4 Cross-encoder classifier . . . 14

4 Experiments and results 14
4.1 Evaluation scores . . . 14
4.2 Bi-encoder retriever . . . 14

5 Conclusion and future works 21
5.1 Conclusion . . . 21
5.2 Future works . . . 22

Refercences 23

List of Figures

1 BERT - the Bidirectional Encoder Representations from Transformers 7
2 Dual Encoder - Dense Passage Retrieval 9
3 Dual Encoder and Cross Encoder 10
4 Architecture of two-stage retriever 11
5 Query Statistic 12
6 Document Statistic 13

List of Tables

1 Results 15

1 Introduction

1.1 Overview

Information retrieval is an important field of Natural Language Processing that have many applications in real life. In this project, we aim to design a reasonable model for the Legal Texts Retrieval Problem - a task in the ZALO AI Competition 2021. The task is to correctly retrieve all relevant legal documents from a legal documents corpus given an input question. The provided datasets include a Vietnamese legal-texts corpus with 61425 documents, a 3196-question-and-answer set for training, and a questions-only set for testing and evaluating. Because answers to the questions-only set are not available, we have to divide the question-and-answer set into 3 sets: train, validation (val), and test for training and evaluating models. Our first idea is to implement a dual dense-encoder, inspired by the famous Dense Passage Retrieval[5] to retrieve relevant articles. However, as the dual encoder itself cannot correctly decide the number of documents retrieved for each question, we have this job done by a cross-encoder. Therefore, we obtain a final two-stage model containing a bi-encoder to select the top k most relevant articles and a cross-encoder to decide which of the k articles are the answers. During the training and experimenting procedure, we made some crucial comparisons to choose out a reasonable final model.

1.2 Problem formulation

This section clearly introduces a challenge of ZALO AI Competition 2021: Legal Text Retrieval task.

1.2.1 Legal documents corpus

ZALO provides competitors with a laws corpus, each law is split into some articles (điều luật) that match to parts of that law. Every article is treated as a legal document and has the following features:

- law id: Identification text of the law to which the article (document) belongs.
- article id: Identification number of the part (of the law) to which the document corresponds. Law id and article id can be the key that identifies a document
- title: Title of the document.
- text: The text in the document.

Here is an example of a document:

- law id: 01/2009/tt-bnn
- article id: 2
- title: Điều 2. Tổ chức lực lượng

- text:

  1. Hàng năm trước mùa mưa, lũ, Ủy ban nhân dân cấp xã nơi có đê phải tổ chức lực lượng lao động tại địa phương để tuần tra, canh gác đê và thường trực trên các điếm canh đê hoặc nhà dân khu vực gần đê (đối với những khu vực chưa có điếm canh đê), khi có báo động lũ từ cấp I trở lên đối với tuyến sông có đê (sau đây gọi tắt là lực lượng tuần tra, canh gác đê).

  2. Lực lượng tuần tra, canh gác đê được tổ chức thành các đội, do Ủy ban nhân dân cấp xã ra quyết định thành lập; từ 01 đến 02 kilômét đê thành lập 01 đội; mỗi đội có từ 12 đến 18 người, trong đó có 01 đội trưởng và 01 hoặc 02 đội phó. Danh sách thành viên đội tuần tra, canh gác đê được niêm yết tại điếm canh đê thuộc địa bàn được phân công.
  3. Khi lũ, bão có diễn biến phức tạp, kéo dài ngày, Uỷ ban nhân dân cấp xã có thể quyết định việc bổ sung thêm thành viên cho đội tuần tra, canh gác đê.

This information is useful in the preprocessing part

1.2.2 Question-Answer set

The task of the challenge can be formulated as below:
Input: A legal question.
Output: All relevant documents that can answer this question. Therefore, in the question and answer set, the information provided is also in that input-output form, the relevant documents are mentioned with their law and article id, while the question is text-written.
Here is an example:
Input: Tử tù chết sau khi tiêm thuốc độc mà không có thân nhân nhận xác thì được hỗ trợ mai táng như thế nào?
Output: relevant articles:

- – law id: 53/2010/qh12

  – article id: 60

  – title: Điều 60. Giải quyết việc xin nhận tử thi, hài cốt của người bị thi hành án tử hình

  – text:

  1. Việc giải quyết nhận tử thi được thực hiện như sau:
  a) Trước khi thi hành án tử hình, thân nhân hoặc người đại diện hợp pháp của người chấp hành án được làm đơn có xác nhận của Ủy ban nhân dân cấp xã nơi cư trú gửi Chánh án Tòa án đã xét xử sơ thẩm đề nghị giải quyết cho nhận tử thi của người chấp hành án để an táng; trường hợp người chấp hành án là người nước ngoài thì đơn phải có xác nhận của cơ quan có thẩm quyền hoặc cơ quan đại diện ngoại giao tại Việt Nam của nước mà người chấp hành án mang quốc tịch và phải được dịch ra tiếng Việt. Đơn phải ghi rõ họ tên, địa chỉ người nhận tử thi, quan hệ với người chấp hành án; cam kết bảo đảm yêu cầu về an ninh, trật tự, vệ sinh môi trường và tự chịu chi phí;
  b) Chánh án Tòa án đã xét xử sơ thẩm thông báo bằng văn bản cho người có đơn

đề nghị về việc cho nhận tử thi hoặc không cho nhận tử thi khi có căn cứ cho rằng việc nhận tử thi ảnh hưởng đến an ninh, trật tự, vệ sinh môi trường. Trường hợp người chấp hành án là người nước ngoài, thì Chánh án Tòa án đã xét xử sơ thẩm có trách nhiệm thông báo bằng văn bản cho Bộ Ngoại giao Việt Nam để thông báo cho cơ quan có thẩm quyền hoặc cơ quan đại diện ngoại giao tại Việt Nam của nước mà người đó mang quốc tịch;
c) Cơ quan thi hành án hình sự Công an cấp tỉnh, cơ quan thi hành án hình sự cấp quân khu có trách nhiệm thông báo cho người có đơn đề nghị ngay sau khi thi hành án để đến nhận tử thi về an táng. Việc giao nhận tử thi phải được thực hiện trong thời hạn 24 giờ kể từ khi thông báo và phải lập biên bản, có chữ ký của các bên giao, nhận; hết thời hạn này mà người có đơn đề nghị không đến nhận tử thi thì cơ quan thi hành án hình sự Công an cấp tỉnh, cơ quan thi hành án hình sự cấp quân khu có trách nhiệm an táng.
2. Trường hợp không được nhận tử thi hoặc thân nhân của người bị thi hành án không có đơn đề nghị được nhận tử thi về an táng thì cơ quan thi hành án hình sự Công an cấp tỉnh, cơ quan thi hành án hình sự cấp quân khu tổ chức việc an táng. Sau 03 năm kể từ ngày thi hành án, thân nhân hoặc đại diện hợp pháp của người đã bị thi hành án được làm đơn có xác nhận của Ủy ban nhân dân cấp xã nơi cư trú đề nghị Cơ quan thi hành án hình sự Công an cấp tỉnh, cơ quan thi hành án hình sự cấp quân khu nơi đã thi hành án cho nhận hài cốt. Đơn đề nghị phải ghi rõ họ tên, địa chỉ người nhận hài cốt, quan hệ với người bị thi hành án; cam kết bảo đảm yêu cầu về an ninh, trật tự, vệ sinh môi trường và tự chịu chi phí. Trong thời hạn 07 ngày, kể từ ngày nhận được đơn, cơ quan thi hành án hình sự Công an cấp tỉnh, cơ quan thi hành án hình sự cấp quân khu có trách nhiệm xem xét, giải quyết.
Trường hợp người bị thi hành án là người nước ngoài thì đơn đề nghị phải có xác nhận của cơ quan có thẩm quyền hoặc cơ quan đại diện ngoại giao tại Việt Nam của nước mà người bị thi hành án mang quốc tịch và phải được dịch ra tiếng Việt. Việc giải quyết cho nhận hài cốt do cơ quan quản lý thi hành án hình sự xem xét, quyết định.

- – law id: 82/2011/nđ-cp
  – article id: 9
  – title: Điều 9. Triển khai việc thi hành án tử hình
  – text:

  1. Căn cứ vào kế hoạch tổ chức thi hành án tử hình của Hội đồng thi hành án tử hình, Cơ quan thi hành án hình sự Công an cấp tỉnh hoặc Cơ quan thi hành án hình sự cấp quân khu có trách nhiệm lập kế hoạch triển khai việc thi hành án tử hình, phân công, bố trí lực lượng, phương tiện cần thiết để đảm bảo cho việc thi hành án tử hình.
  2. Chủ tịch Hội đồng thi hành án tử hình ra quyết định hoặc có văn bản yêu cầu Sở Y tế tỉnh, thành phố trực thuộc Trung ương hoặc Phòng Quân y cấp quân khu nơi Tòa án đã ra quyết định thi hành án cử bác sỹ của bệnh viện thuộc Sở Y tế hoặc bệnh viện thuộc quân khu đến địa điểm thi hành án tử hình để hỗ trợ việc xác định

tĩnh mạch của người bị thi hành án tử hình trong trường hợp cần thiết.
3. Trường hợp người bị thi hành án tử hình là phụ nữ thì ngay sau khi nhận đủ hồ sơ để đưa bản án tử hình ra thi hành, Hội đồng thi hành án tử hình phải yêu cầu Thủ trưởng Cơ quan thi hành án hình sự Công an cấp tỉnh hoặc Thủ trưởng Cơ quan thi hành án hình sự cấp quân khu ra lệnh trích xuất người bị kết án tử hình đến bệnh viện thuộc Sở Y tế hoặc bệnh viện thuộc quân khu nơi đang giam giữ người bị kết án tử hình để kiểm tra, xác định xem người bị kết án tử hình có thai hay không. Việc kiểm tra phải được lập thành văn bản và có xác nhận của bệnh viện nơi tiến hành kiểm tra, xác định.

Of all 3196 questions in the set:

- Most of them (3103 questions) only have one relevant articles.
- There are 84 queries having two relevant documents and only 9 having three.

We split these 3196 queries into three sets train: validation: test following the ratio 2400: 350: 446. The split data is used later for training and evaluating models.

After filtering identical documents, we received total 60830 articles from 3263 legal documents.

2 Theoretical preliminaries

2.1 BM25+

2.1.1 BM25

BM25 (BM stands for best matching) bag-of-word ranking function is used by search engines to determine how relevant documents are to a particular search query based on the query terms appearing in each document, regardless of their proximity within the document. This method is based on a probabilistic retrieval framework developed in the 1970s and 1980s by Stephen E. Robertson, Karen Spärck Jones, and others. The BM25 and its newer variants, e.g. BM25+, are retrieval functions similar to TF-IDF.

Suppose we have a query Q, containing keywords $q_1, q_2, ..., q_n$, the formula to calculate BM25 score of a document D is as follows:

$$\textbf{score}(D,Q) = \sum_{i=0}^{n} \textbf{IDF}(q_i).\frac{f(q_i,D).(k_1+1)}{f(q_i,D)+k_1.(1-b+b.\frac{|D|}{avgdl})}$$

where $f(q_i, D)$ is the number of times that q_i occurs in D, $|D|$ is the number of words (length) of D and $avgdl$ is the average length of documents in the corpus. k_1 and b are tuning parameters, usually chosen as $k_1 \in [1.2, 2.0]$ and $b = 0.75$. $\textbf{IDF}(q_i)$ is the inverse document frequency (IDF) weight of the word q_i, calculated as:

$$\textbf{IDF}(q_i) = ln(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5})$$

where N is the number of documents in the corpus, and $n(q_i)$ is the number of documents containing q_i.

2.1.2 BM25+

This variant of BM25 has two significant changes like below:

$$\textbf{score}(D,Q) = \sum_{i=0}^{n} ln(\frac{N+1}{n(q_i)+0.5}).(\frac{f(q_i,D).(k_1+1)}{f(q_i,D)+k_1.(1-b+b.\frac{|D|}{avgdl})} + \delta)$$

The first difference is that in BM25+, the IDF weight is changed simpler to avoid negative values when the number of documents containing a word is more than half the number of documents in the corpus. The second one is the appearance of another parameter δ. This was created to address a flaw in the standard BM25 in which the component of term frequency normalization by document length is not properly lower-bounded; as a result of this flaw, long documents that do match the query term are frequently scored unfairly by BM25 as having a similar relevancy to shorter documents that do not contain the query term at all [7]. The authors of this method concluded that with $\delta = 1$, BM25+ can work well with all corpus and outperforms the original BM25.

In the project, BM25+ is used to generate hard-negative documents in the training procedure of the bi-encoder model, which is described later. I utilized the $rank - bm25$ library which follows strictly the above formula from the paper Improvements to BM25 and Language Models Examined[10].

2.2 PhoBERT

BERT-the Bidirectional Encoder Representations from Transformers[1] has become extremely popular over the past few years and has contributed to significant improvement gains in various aspects of natural language processing (NLP). It represents word tokens by embedding vectors that encode the contexts where the words appear, i.e. contextualized word embeddings. Moreover, with a fine-tuning procedure on specified tasks, we can get sentence, and paragraph embeddings from BERT, which can be served later in classification, ranking... However, the success of pre-trained BERT and its variants has largely been limited to the English language. For other languages, one could retrain a language-specific model using the BERT architecture or employ existing pre-trained multilingual BERT-based models.

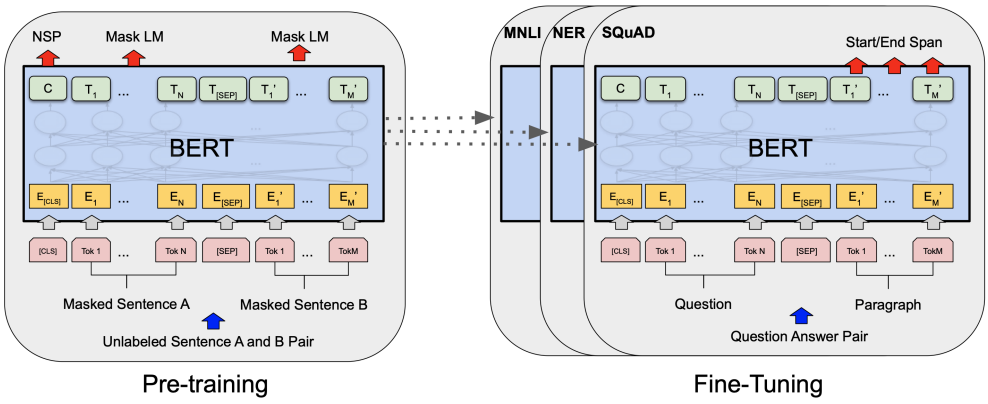

Figure 1: BERT - the Bidirectional Encoder Representations from Transformers

In the field of Vietnamese NLP, several publicly released monolingual and multilingual BERT models fail to detect the difference between Vietnamese syllables and word tokens and are pre-trained with syllable-level Vietnamese text data. This concern may lead to poor performance in some tasks that require models pre-trained on word-level data.

PhoBERT[8], publicly released in 2020 by VinAI researchers, is a model that can solve the above problem. The model is trained on a large word-level Vietnamese corpus following the robust training procedure of RoBERTa[6]. It is not a surprise that PhoBERT immediately brings state-of-the-art results on many Vietnamese NLP tasks, such as NER, NLI, POS-tagging... Therefore, in this project, I utilize PhoBERT in both the bi-encoder and cross-encoder, which is described later. The implementation uses the public $PhoBERT_{base}$ model from $transformer$ library, along with the segmenter from $vncorenlp$[11] to get the word-level data.

2.3 Dual Encoder and Cross Encoder in information retrieval

In the field of Information Retrieval, one crucial mission is to learn the relationship between a query and documents. Some traditional methods like TF-IDF, BM25 (BM25+) can only learn the similarity of word appearance and ignore the latent semantic. Therefore, these models do not have the ability to detect synonyms and paraphrases which have different tokens but close meanings. That is the main reason that leads to the inefficiency of these methods, especially in some kind of corpus.

To handle this problem, the idea of learning relationships through token (word) dense embedding vectors with deep models has been proposed and become more and more popular in recent years. With the appearance of BERT and its variants which can represent tokens efficiently, these methods are developed fast and produce better and better results, outperforming traditional ones in many areas. In this project, we use two important techniques in this field.

2.3.1 Dual Encoder (Bi-Encoder): Dense Passage Retrieval[5]

This method combines two encoders used for representing the query and document and then learns the relationship between the embedding vectors. The encoders usually have the same architecture; however, after the training or fine-tuning procedures, they will have different parameters that are suitable for their different tasks of embedding queries and documents. The authors of Dense Passage Retrieval have introduced the idea of using two BERT models to obtain the dense vectors, which is the final representation of the [CLS] token. The relationship is decided through the inner product or cosine similarity of the vectors. This method is shown to have good performance despite not having added expensive pre-training procedures, for instance, inverse cloze task (ICT). In the paper, the authors show that fine-tuning the two models on a small set of labeled question-document datasets is enough, as long as the procedure uses hard-negative documents - articles that do not contain answers for the question (query) but have close meaning. The hard negatives of each query are obtained by BM25. The implementation of the bi-encoder is fast and easy with the $faiss$[4] library, which support fast similarity search on large-scale data. One problem of the dual encoder is that it can only determine the relevant relationship through inner-product scores. Therefore, the bi-encoder can only be applied to retrieve top k (k is a fixed integer) relevant articles from the corpus given a query. The k documents need to be determined further whether they are the answers or not.

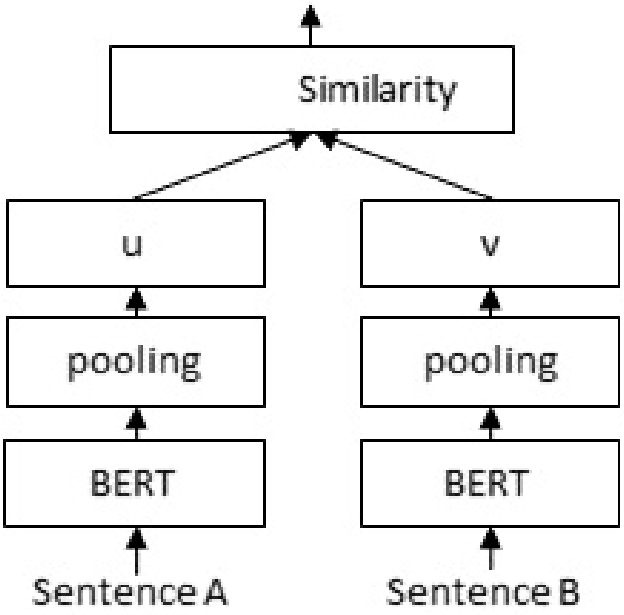

Figure 2: Dual Encoder - Dense Passage Retrieval

In the original paper Dense Passage Retrieval, the authors use the inner product to determine the relevance. The aim of the training procedure is to create good encoders that match questions and articles into a vector space where relevant query-document are close to each other, which means their vector inner product is high.

Let $P = (< q_i, d_i^+, d_{i,1}^-, ..., d_{i,n}^- >)_{i=1}^m$ is the training data having m queries. Each question q_i has one positive (relevant) document d_i^+ and n negative (irrelevant) documents $p_{i,j}^-$. The job is to optimize the negative log-likelihood loss function of the positive document:

$$\mathbf{L}(q_i, d_i^+, d_{i,1}^-, ..., d_{i,n}^-) = -log(\frac{e^{q_i.d_i^+}}{e^{q_i.d_i^+} + \sum_{j=1}^n e^{q_i.d_{i,j}^-}})$$

In practice, the training procedure can only handle this loss in a batch containing a small number of questions, each having a relevant document. Positive documents of other queries are considered to be the random in batch negative documents. Adding a hard-negative document with each query obtained by BM25 can increase the model performance; however, the performance is not always better if we increase the number of hard-negative documents.

2.3.2 Cross-Encoder classifier

Unlike Bi-encoder, Cross-encoder does not learn to embed queries and documents. It concatenates the query and article into one input and sends it through a deep model with a classification layer, at last, to decide whether they are relevant or not. Using BERT variants, the encoder can learn the relationship between tokens of the query and documents. In other words, the cross-encoder learns an embedding vector of the concatenation (which is chosen as the representation of the [CLS] token in the last layer when using the BERT-variant-encoder) and then puts it

into a classification layer. Therefore, I can use it to decide whether a document is relevant to the query or not. However, it is expensive and time-consuming to concatenate each query with every article in the corpus; therefore, the encoder needs a previous-stage filter, which can be BM25 or bi-encoder to quickly choose out k good articles. The training procedure can use the Cross-Entropy loss.

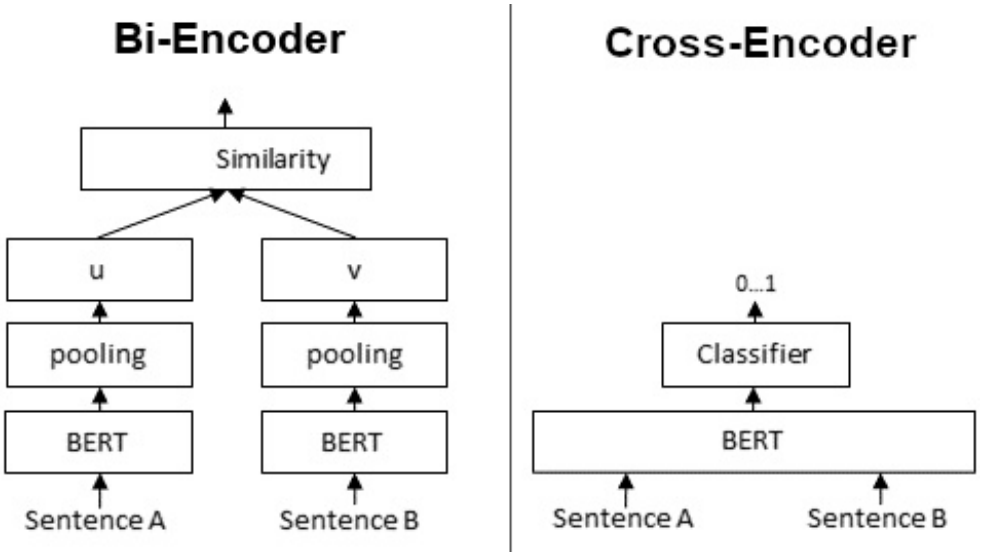

Figure 3: Dual Encoder and Cross Encoder

3 The two-stage model for legal text retrieval problem

This chapter described fully the architecture of our retrieval model.

3.1 Two-stage retrieval system

With the theory introduced in the previous chapter, we have our system detailed below:

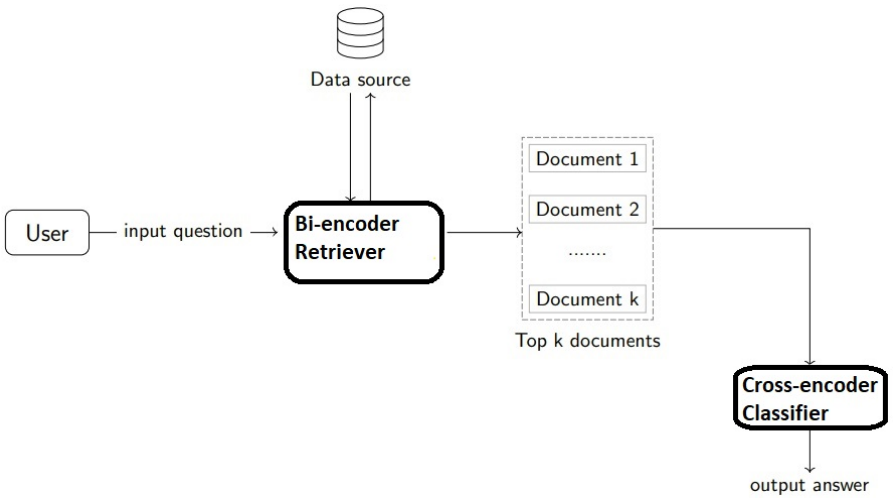

Figure 4: Architecture of two-stage retriever

- An input question provided by the user is sent to the system. It is processed (lower, remove stop-words or end-phrases, segment...).
- The processed query is fed to the query-encoder of the bi-encoder retriever to obtain a query embedding vector. The system then quickly computes and compares the inner products of the query embedding and all document embeddings, which are pre-saved. Top k documents (k is an integer) with the highest inner-product scores are picked and their processed versions are concatenated with the processed query.
- k concatenations is put into the cross-encoder classification. The system returns documents that have the corresponding concatenations classified as relevant. If there is no document labeled as relevant, the one having the highest inner product score with the query embedding is returned.

The initialization of the model requires the document-corpus, a trained bi-encoder retriever and a trained cross-encoder classifier.

3.2 Processing technique

3.2.1 PhoBERT processing

As mentioned, the retrieval system implements phoBERT in the dual encoder and cross encoder; therefore, the queries, documents and their concatenation both need processing.

- Query:
  - Lower every character except for abbreviations (UBND, BHYT...).
  - Remove legal stop words (chương...,ngày...tháng...năm, thông tư, nghị quyết...).
  - Remove punctuations and only keeps . and ,.

  - Remove question end-phrases (là gì, như nào, thế nào, ...).
  - Segment into word-level with *vncorenlp* library.

- Documents:
  - Put the document title in the head of the document text. However remove Điều... phrases.
  - Lower every character excepts for abbreviations.
  - Remove newline character, indexes and bad characters.
  - Remove legal stop words (chương...,ngày...tháng...năm, thông tư, nghị quyết...)
  - Remove punctuations and only keeps . and ,.
  - Segment into word-level with *pyvi* library.

- Query-Document concatenation: Concatenate processed query and processed document with the [SEP] (</s>) token in the middle.

After processing, we figure some characteristics of the query and document lengths in words:

	ques_len
count	3196.000000
mean	10.598561
std	3.941352
min	1.000000
25%	8.000000
50%	10.000000
75%	13.000000
max	28.000000

Figure 5: Query Statistic

Of all 3196 questions after processing, the one with 28 segmented words is the longest. Most of the questions are below 13 words. Therefore, we use 32 as the padding length in phoBERT query encoder

	id	len
count	61425.000000	61425.000000
mean	30712.000000	209.298722
std	17732.014479	294.809171
min	0.000000	2.000000
25%	15356.000000	74.000000
50%	30712.000000	136.000000
75%	46068.000000	251.000000
max	61424.000000	12089.000000

Figure 6: Document Statistic

The corpus has a total 61425 documents and 75 percent of them have lengths smaller or equal to 251. However, there are some documents that are extremely long. The maximum input length of PhoBERT tokenizer is 256 and we decided to choose it as the padding length in phoBERT document encoder. We cannot cut long documents into smaller passages because this requires relabeled data by hand, which violates the regulations competition.

Similar to the documents, we use padding length 256 for all concatenations in phoBERT cross-encoder.

3.2.2 BM25+ processing

As the training procedure requires using hard-negative documents obtained with BM25+, we also need another processing stage on the query and documents, which is quite the same as phoBERT processing but does not have the word-segment step. Moreover, every punctuation should be removed.

3.3 Bi-encoder retriever

In the project, we implement the dense passage retrieval with a $\boldsymbol{PhoBERT_{base}}$ query encoder and a $\boldsymbol{PhoBERT_{base}}$ document encoder, obtain hard-negative documents by BM25+, learn query-document relationship through inner product, handle indexes with $faiss$ library but have some important changes in the training procedure. The dual-encoder plays the role of a filter that chooses out k promising documents for further determining in the cross-encoder.

3.3.1 Number of hard negative documents

As our training data is small, for each training query, we obtain multiple hard-negative documents using BM25+. In this project, we experiment with a number of hard negative documents (n_{hard}) of 0, 1, 3 and 7.

3.3.2 Finetuning before training

As the data is domain-specific, we experiment with fine-tuning before training dual-encoder to strengthen PhoBERT encoder with the ability to understand the legal context. We try masked language model (MLM) [1] and Condenser [2] approaches. Building the dataset for training bi-encoder includes grouping a query with only a positive document and n hard-negative documents obtained by BM25+. Therefore, with 2400 divided training questions, we obtained 2483 training groups. After each training epoch, the evaluation is only conducted in batches on same-built validation data as testing on the whole corpus is expensive. Therefore, the evaluation step in training bi-encoder is not important.

3.3.3 Training techniques

We use the technique to extend batch size despite limited GPU memories [3] and the two stage training inspired from [9], which show their effectiveness.

3.4 Cross-encoder classifier

The implementing cross-encoder in the systems using phoBERT and Cross-Entropy loss for training coding with sentencetransformers framework. To build the train, validation and test data, we use the trained bi-encoder model to generate top 30 relevant documents for each query. In those 30 articles, the ones that are not labeled to be positive are concatenated with the query to be negative examples. To scale the proportion of negatives and positives, we duplicate and positive query-document concatenation 29 times, as almost all queries only have one relevant document. During the training procedure, the epoch state that shows the best performance on the validation set is saved to build the final retrieval system.

4 Experiments and results

This part shows training results of the bi-encoder and cross-encoder as well as the performance of final retrieval model. All training procedures are conducted on Kaggle with GPU NVIDIA Tesla P100.

4.1 Evaluation scores

In this project, we use hit accuracy score for evaluating models, that is the proportion of samples having at least one positive passage returned in the top k retrieved.

4.2 Bi-encoder retriever

For the bi-encoder, we mainly follow the training setup of the original paper Dense Passage Retrieval that has a learning rate of $10^{-}5$ using Adam, linear scheduling with warm-up and dropout rate 10^{-5}. However, batch size and number of training epoch are reduced due to limited resources. For model using n, we mainly train with batch size 64,128,256 and 40 training epochs in first stage and 15 in the second stage. To build the bi-encoder retriever, trained document

encoder is then used to create *faiss* index for all the corpus, and trained query encoder is used to embed questions for retrieving articles. Results are obtained with top 1,5,10,30,100 documents on test data, evaluated with hit scores. Conducting experiments, we make some important conclusions:

- Increasing the batch size and number of hard leads to the better results.
- MLM is the best starting model. Although the Condenser is shown to be better in [2], it might not be optimal for this problem. As the parameters of added layers are just initialized, the model needs huge training data and a long training procedure to outperform the model pretrained with only masked language modelling.
- Training the second stage does improve the results of bi-encoder. As the hard negatives are updated with the recently-trained state from the first stage, the final model obtained from the second stage may have the ability to distinguish both lexical and semantic hard negatives.
- Chunking passages helps the system.
- Cross-encoder is a good re-ranker. However, in the future, we will try to combining the scores of both dual-encoder and cross-encoder to strengthen the re-ranker.

stage	model/setting	batch	n	Top 1	Top 5	Top 10	Top 30	Top 100
1	PhoBERT-base-v2	64	0	54.26	83.63	87.44	92.60	95.51
	PhoBERT-base-v2	64	1	62.78	84.52	89.01	93.27	96.86
	PhoBERT-base-v2	64	3	67.93	86.77	90.35	94.39	96.86
	PhoBERT-base-v2	64	7	67.04	85.87	90.58	94.17	97.31
	PhoBERT-base-v2	128	7	67.04	86.55	90.13	95.07	97.53
	PhoBERT-base-v2	256	7	68.83	87.44	90.81	95.07	97.53
	Masked LM	256	7	70.85	89.69	93.27	96.64	98.21
	Condenser	256	7	69.73	88.57	91.03	95.74	97.98
	MLM + Condenser	256	7	70.18	89.01	92.15	96.41	97.76
2	all articles	256	7	74.22	91.26	93.95	96.41	98.21
	chunks			76.91	91.03	94.61	97.09	98.21
re-ranked	chunks	-	-	84.52	94.61	96.41	97.09	-

Table 1: Results

Here are some common mistakes of the system when $k = 30$:

- Retrieve wrong answers. Example:

  Input: Lãi suất công cụ nợ của Chính phủ được quy định như thế nào?

  Answer(s): Điều 19. Lãi suất
  Lãi suất trong giao dịch mua bán lại, giao dịch vay và cho vay phù hợp với quy định pháp luật khác, được tính trên cơ sở ngày thực tế/ngày thực tế. Cách tính lãi suất được quy

định cụ thể trong quy định nghiệp vụ của Sở Giao dịch Chứng khoán hoặc của Trung tâm Lưu ký Chứng khoán Việt Nam (đối với loại hình giao dịch vay và cho vay tại Trung tâm Lưu ký Chứng khoán Việt Nam).

Retrieved Article(s): Điều 12. Lãi suất
1. Ngân hàng Nhà nước công bố lãi suất tái cấp vốn, lãi suất cơ bản và các loại lãi suất khác để điều hành chính sách tiền tệ, chống cho vay nặng lãi.
2. Trong trường hợp thị trường tiền tệ có diễn biến bất thường, Ngân hàng Nhà nước quy định cơ chế điều hành lãi suất áp dụng trong quan hệ giữa các tổ chức tín dụng với nhau và với khách hàng, các quan hệ tín dụng khác.

Reason: Retrieved article has quite similar words as answer. The model can not distinguish the meaning of two documents.

- Retrieve not enough correct answers

  Input: Không đăng ký tạm trú cho khách nước ngoài phạt bao nhiêu tiền?

  Answer(s):

  – Điều 17. Vi phạm các quy định về xuất cảnh, nhập cảnh, quá cảnh, cư trú và đi lại
    1. Phạt cảnh cáo hoặc phạt tiền từ 100.000 đồng đến 300.000 đồng đối với người nước ngoài đi lại trên lãnh thổ Việt Nam mà không mang theo hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu.
    2. Phạt tiền từ 500.000 đồng đến 2.000.000 đồng đối với một trong những hành vi sau đây:
    a) Không thông báo ngay cho cơ quan có thẩm quyền về việc mất, hư hỏng hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu, thị thực Việt Nam, thẻ tạm trú, thẻ thường trú;
    b) Tẩy, xóa, sửa chữa hoặc làm sai lệch hình thức, nội dung ghi trong hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu, thị thực, thẻ tạm trú và thẻ thường trú;
    c) Khai không đúng sự thật để được cấp hộ chiếu, giấy tờ khác có giá trị thay hộ chiếu, thị thực Việt Nam, thẻ tạm trú, thẻ thường trú hoặc giấy tờ có giá trị nhập cảnh, xuất cảnh, cư trú tại Việt Nam;
    d) Người nước ngoài đi vào khu vực cấm, khu vực nhà nước quy định cần có giấy phép mà không có giấy phép hoặc đi lại quá phạm vi, thời hạn được phép;
    đ) Không xuất trình hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu hoặc giấy tờ có liên quan đến xuất nhập cảnh khi nhà chức trách Việt Nam yêu cầu; không chấp hành các yêu cầu khác của nhà chức trách Việt Nam về kiểm tra người, hành lý;
    e) Người nước ngoài không khai báo tạm trú theo quy định hoặc sử dụng chứng nhận tạm trú, thẻ tạm trú, thẻ thường trú ở Việt Nam quá thời hạn từ 15 ngày trở xuống mà không được cơ quan có thẩm quyền cho phép;
    g) Cho người nước ngoài nghỉ qua đêm nhưng không khai báo tạm trú, không hướng dẫn người nước ngoài khai báo tạm trú theo quy định hoặc không thực hiện đúng các quy định khác của cơ quan có thẩm quyền.
    3. Phạt tiền từ 3.000.000 đồng đến 5.000.000 đồng đối với một trong những hành vi sau đây:
    a) Qua lại biên giới quốc gia mà không làm thủ tục xuất cảnh, nhập cảnh theo quy

định;
b) Trốn hoặc tổ chức, giúp đỡ người khác trốn vào các phương tiện nhập cảnh, xuất cảnh nhằm mục đích vào Việt Nam hoặc ra nước ngoài;
c) Cho người khác sử dụng hộ chiếu, giấy tờ có giá trị thay hộ chiếu để thực hiện hành vi trái quy định của pháp luật;
d) Sử dụng hộ chiếu hoặc các giấy tờ khác có giá trị thay hộ chiếu của người khác để nhập cảnh, xuất cảnh, quá cảnh;
đ) Người nước ngoài không khai báo tạm trú theo quy định hoặc sử dụng chứng nhận tạm trú, thẻ tạm trú, thẻ thường trú ở Việt Nam quá thời hạn từ 16 ngày trở lên mà không được cơ quan có thẩm quyền cho phép;
e) Người nước ngoài đã được cấp thẻ thường trú mà thay đổi địa chỉ nhưng không khai báo để thực hiện việc cấp đổi lại.
4. Phạt tiền từ 5.000.000 đồng đến 10.000.000 đồng đối với một trong những hành vi sau đây:
a) Chủ phương tiện, người điều khiển các loại phương tiện chuyên chở người nhập cảnh, xuất cảnh Việt Nam trái phép;
b) Sử dụng hộ chiếu giả, giấy tờ có giá trị thay hộ chiếu giả, thị thực giả, thẻ tạm trú giả, thẻ thường trú giả, dấu kiểm chứng giả để xuất cảnh, nhập cảnh, quá cảnh, cư trú.
5. Phạt tiền từ 15.000.000 đồng đến 25.000.000 đồng đối với một trong những hành vi sau đây:
a) Giúp đỡ, chứa chấp, che giấu, tạo điều kiện cho người khác đi nước ngoài, ở lại nước ngoài, vào Việt Nam, ở lại Việt Nam hoặc qua lại biên giới quốc gia trái phép;
b) Người nước ngoài nhập cảnh, hành nghề hoặc có hoạt động khác tại Việt Nam mà không được phép của cơ quan có thẩm quyền của Việt Nam;
c) Cá nhân, tổ chức ở Việt Nam bảo lãnh hoặc làm thủ tục cho người nước ngoài nhập cảnh Việt Nam, xin cấp thị thực, cấp thẻ tạm trú, gia hạn tạm trú, giấy tờ có giá trị nhập cảnh, cư trú tại Việt Nam nhưng không thực hiện đúng trách nhiệm theo quy định của pháp luật hoặc khai không đúng sự thật khi bảo lãnh, mời hoặc làm thủ tục cho người nước ngoài nhập cảnh, xin cấp thị thực, cấp thẻ tạm trú, gia hạn tạm trú, giấy tờ có giá trị nhập cảnh, xuất cảnh, cư trú tại Việt Nam;
d) Người nước ngoài nhập cảnh hoạt động không đúng mục đích, chương trình đã đề nghị xin cấp thẻ tạm trú, thẻ thường trú.
6. Phạt tiền từ 30.000.000 đồng đến 40.000.000 đồng đối với một trong những hành vi sau đây:
a) Giả mạo hồ sơ, giấy tờ để được cấp hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu, thị thực, thẻ tạm trú, thẻ thường trú;
b) Làm giả hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu, thị thực, thẻ tạm trú, thẻ thường trú hoặc dấu kiểm chứng;
c) Trốn vào đại sứ quán, lãnh sự quán hoặc trụ sở cơ quan, tổ chức quốc tế đóng tại Việt Nam;
d) Người nước ngoài cư trú tại Việt Nam mà không được phép của cơ quan có thẩm quyền;
đ) Tổ chức, đưa dẫn hoặc môi giới cho người khác xuất cảnh, nhập cảnh Việt Nam trái phép.

7. Hình thức xử phạt bổ sung:
Tịch thu tang vật, phương tiện vi phạm hành chính đối với hành vi quy định tại Điểm b Khoản 2; Điểm c, d Khoản 3; Điểm a Khoản 4; Điểm a, b Khoản 6 Điều này.
8. Biện pháp khắc phục hậu quả:
a) Buộc thu hồi hộ chiếu, giấy tờ khác có giá trị thay hộ chiếu, thị thực, thẻ tạm trú, thẻ thường trú hoặc dấu kiểm chứng đối với hành vi quy định tại Điểm b Khoản 2; Điểm d, đ Khoản 3; Điểm b Khoản 4; Điểm a, b Khoản 6 Điều này;
b) Buộc hủy bỏ thông tin, tài liệu sai sự thật đối với hành vi quy định tại Điểm c Khoản 2; Điểm c Khoản 5 Điều này.
9. Người nước ngoài có hành vi vi phạm hành chính quy định tại Khoản 1, 2, 3, 4, 5 và Khoản 6 Điều này, thì tùy theo mức độ vi phạm có thể bị áp dụng hình thức xử phạt trục xuất khỏi nước Cộng hòa xã hội chủ nghĩa Việt Nam

– Điều 33. Khai báo tạm trú
1. Người nước ngoài tạm trú tại Việt Nam phải thông qua người trực tiếp quản lý, điều hành hoạt động của cơ sở lưu trú để khai báo tạm trú với Công an xã, phường, thị trấn hoặc đồn, trạm Công an nơi có cơ sở lưu trú.
2. Người trực tiếp quản lý, điều hành hoạt động của cơ sở lưu trú có trách nhiệm ghi đầy đủ nội dung mẫu phiếu khai báo tạm trú cho người nước ngoài và chuyển đến Công an xã, phường, thị trấn hoặc đồn, trạm Công an nơi có cơ sở lưu trú trong thời hạn 12 giờ, đối với địa bàn vùng sâu, vùng xa trong thời hạn là 24 giờ kể từ khi người nước ngoài đến cơ sở lưu trú.
3. Cơ sở lưu trú du lịch là khách sạn phải nối mạng Internet hoặc mạng máy tính với cơ quan quản lý xuất nhập cảnh Công an tỉnh, thành phố trực thuộc trung ương để truyền thông tin khai báo tạm trú của người nước ngoài. Cơ sở lưu trú khác có mạng Internet có thể gửi trực tiếp thông tin khai báo tạm trú của người nước ngoài theo hộp thư điện tử công khai của cơ quan quản lý xuất nhập cảnh Công an tỉnh, thành phố trực thuộc trung ương.
4. Người nước ngoài thay đổi nơi tạm trú hoặc tạm trú ngoài địa chỉ ghi trong thẻ thường trú thì phải khai báo tạm trú theo quy định tại khoản 1 Điều này.

Retrieved Article(s):

– Điều 17. Vi phạm các quy định về xuất cảnh, nhập cảnh, quá cảnh, cư trú và đi lại
1. Phạt cảnh cáo hoặc phạt tiền từ 100.000 đồng đến 300.000 đồng đối với người nước ngoài đi lại trên lãnh thổ Việt Nam mà không mang theo hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu.
2. Phạt tiền từ 500.000 đồng đến 2.000.000 đồng đối với một trong những hành vi sau đây:
a) Không thông báo ngay cho cơ quan có thẩm quyền về việc mất, hư hỏng hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu, thị thực Việt Nam, thẻ tạm trú, thẻ thường trú;
b) Tẩy, xóa, sửa chữa hoặc làm sai lệch hình thức, nội dung ghi trong hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu, thị thực, thẻ tạm trú và thẻ thường trú;
c) Khai không đúng sự thật để được cấp hộ chiếu, giấy tờ khác có giá trị thay hộ

chiếu, thị thực Việt Nam, thẻ tạm trú, thẻ thường trú hoặc giấy tờ có giá trị nhập cảnh, xuất cảnh, cư trú tại Việt Nam;
d) Người nước ngoài đi vào khu vực cấm, khu vực nhà nước quy định cần có giấy phép mà không có giấy phép hoặc đi lại quá phạm vi, thời hạn được phép;
đ) Không xuất trình hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu hoặc giấy tờ có liên quan đến xuất nhập cảnh khi nhà chức trách Việt Nam yêu cầu; không chấp hành các yêu cầu khác của nhà chức trách Việt Nam về kiểm tra người, hành lý;
e) Người nước ngoài không khai báo tạm trú theo quy định hoặc sử dụng chứng nhận tạm trú, thẻ tạm trú, thẻ thường trú ở Việt Nam quá thời hạn từ 15 ngày trở xuống mà không được cơ quan có thẩm quyền cho phép;
g) Cho người nước ngoài nghỉ qua đêm nhưng không khai báo tạm trú, không hướng dẫn người nước ngoài khai báo tạm trú theo quy định hoặc không thực hiện đúng các quy định khác của cơ quan có thẩm quyền.
3. Phạt tiền từ 3.000.000 đồng đến 5.000.000 đồng đối với một trong những hành vi sau đây:
a) Qua lại biên giới quốc gia mà không làm thủ tục xuất cảnh, nhập cảnh theo quy định;
b) Trốn hoặc tổ chức, giúp đỡ người khác trốn vào các phương tiện nhập cảnh, xuất cảnh nhằm mục đích vào Việt Nam hoặc ra nước ngoài;
c) Cho người khác sử dụng hộ chiếu, giấy tờ có giá trị thay hộ chiếu để thực hiện hành vi trái quy định của pháp luật;
d) Sử dụng hộ chiếu hoặc các giấy tờ khác có giá trị thay hộ chiếu của người khác để nhập cảnh, xuất cảnh, quá cảnh;
đ) Người nước ngoài không khai báo tạm trú theo quy định hoặc sử dụng chứng nhận tạm trú, thẻ tạm trú, thẻ thường trú ở Việt Nam quá thời hạn từ 16 ngày trở lên mà không được cơ quan có thẩm quyền cho phép;
e) Người nước ngoài đã được cấp thẻ thường trú mà thay đổi địa chỉ nhưng không khai báo để thực hiện việc cấp đổi lại.
4. Phạt tiền từ 5.000.000 đồng đến 10.000.000 đồng đối với một trong những hành vi sau đây:
a) Chủ phương tiện, người điều khiển các loại phương tiện chuyên chở người nhập cảnh, xuất cảnh Việt Nam trái phép;
b) Sử dụng hộ chiếu giả, giấy tờ có giá trị thay hộ chiếu giả, thị thực giả, thẻ tạm trú giả, thẻ thường trú giả, dấu kiểm chứng giả để xuất cảnh, nhập cảnh, quá cảnh, cư trú.
5. Phạt tiền từ 15.000.000 đồng đến 25.000.000 đồng đối với một trong những hành vi sau đây:
a) Giúp đỡ, chứa chấp, che giấu, tạo điều kiện cho người khác đi nước ngoài, ở lại nước ngoài, vào Việt Nam, ở lại Việt Nam hoặc qua lại biên giới quốc gia trái phép;
b) Người nước ngoài nhập cảnh, hành nghề hoặc có hoạt động khác tại Việt Nam mà không được phép của cơ quan có thẩm quyền của Việt Nam;
c) Cá nhân, tổ chức ở Việt Nam bảo lãnh hoặc làm thủ tục cho người nước ngoài nhập cảnh Việt Nam, xin cấp thị thực, cấp thẻ tạm trú, gia hạn tạm trú, giấy tờ có giá trị nhập cảnh, cư trú tại Việt Nam nhưng không thực hiện đúng trách nhiệm theo quy định của pháp luật hoặc khai không đúng sự thật khi bảo lãnh, mời hoặc

làm thủ tục cho người nước ngoài nhập cảnh, xin cấp thị thực, cấp thẻ tạm trú, gia hạn tạm trú, giấy tờ có giá trị nhập cảnh, xuất cảnh, cư trú tại Việt Nam;
d) Người nước ngoài nhập cảnh hoạt động không đúng mục đích, chương trình đã đề nghị xin cấp thẻ tạm trú, thẻ thường trú.
6. Phạt tiền từ 30.000.000 đồng đến 40.000.000 đồng đối với một trong những hành vi sau đây:
a) Giả mạo hồ sơ, giấy tờ để được cấp hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu, thị thực, thẻ tạm trú, thẻ thường trú;
b) Làm giả hộ chiếu hoặc giấy tờ khác có giá trị thay hộ chiếu, thị thực, thẻ tạm trú, thẻ thường trú hoặc dấu kiểm chứng;
c) Trốn vào đại sứ quán, lãnh sự quán hoặc trụ sở cơ quan, tổ chức quốc tế đóng tại Việt Nam;
d) Người nước ngoài cư trú tại Việt Nam mà không được phép của cơ quan có thẩm quyền;
đ) Tổ chức, đưa dẫn hoặc môi giới cho người khác xuất cảnh, nhập cảnh Việt Nam trái phép.
7. Hình thức xử phạt bổ sung:
Tịch thu tang vật, phương tiện vi phạm hành chính đối với hành vi quy định tại Điểm b Khoản 2; Điểm c, d Khoản 3; Điểm a Khoản 4; Điểm a, b Khoản 6 Điều này.
8. Biện pháp khắc phục hậu quả:
a) Buộc thu hồi hộ chiếu, giấy tờ khác có giá trị thay hộ chiếu, thị thực, thẻ tạm trú, thẻ thường trú hoặc dấu kiểm chứng đối với hành vi quy định tại Điểm b Khoản 2; Điểm d, đ Khoản 3; Điểm b Khoản 4; Điểm a, b Khoản 6 Điều này;
b) Buộc hủy bỏ thông tin, tài liệu sai sự thật đối với hành vi quy định tại Điểm c Khoản 2; Điểm c Khoản 5 Điều này.
9. Người nước ngoài có hành vi vi phạm hành chính quy định tại Khoản 1, 2, 3, 4, 5 và Khoản 6 Điều này, thì tùy theo mức độ vi phạm có thể bị áp dụng hình thức xử phạt trục xuất khỏi nước Cộng hòa xã hội chủ nghĩa Việt Nam

– Điều 8. Vi phạm quy định về đăng ký và quản lý cư trú
1. Phạt tiền từ 100.000 đồng đến 300.000 đồng đối với một trong những hành vi sau đây:
a) Cá nhân, chủ hộ gia đình không thực hiện đúng quy định về đăng ký thường trú, đăng ký tạm trú hoặc điều chỉnh những thay đổi trong sổ hộ khẩu, sổ tạm trú;
b) Cá nhân, chủ hộ gia đình không thực hiện đúng quy định về thông báo lưu trú, khai báo tạm vắng;
c) Không chấp hành việc kiểm tra hộ khẩu, kiểm tra tạm trú, kiểm tra lưu trú hoặc không xuất trình sổ hộ khẩu, sổ tạm trú, giấy tờ khác liên quan đến cư trú theo yêu cầu của cơ quan có thẩm quyền.
2. Phạt tiền từ 1.000.000 đồng đến 2.000.000 đồng đối với một trong những hành vi sau đây:
a) Tẩy, xóa, sửa chữa hoặc có hành vi khác làm sai lệch nội dung sổ hộ khẩu, sổ tạm trú, giấy tờ khác liên quan đến cư trú;
b) Cung cấp thông tin, tài liệu sai sự thật về cư trú;
c) Thuê, cho thuê sổ hộ khẩu, sổ tạm trú, giấy tờ khác liên quan đến cư trú để thực

hiện hành vi trái quy định của pháp luật;
d) Sử dụng sổ hộ khẩu, sổ tạm trú, giấy tờ khác liên quan đến cư trú để thực hiện hành vi trái quy định của pháp luật;
đ) Cơ sở kinh doanh lưu trú không thực hiện việc thông báo lưu trú với cơ quan công an theo quy định khi có người đến lưu trú;
e) Tổ chức kích động, xúi giục, lôi kéo, dụ dỗ, môi giới, cưỡng bức người khác vi phạm pháp luật về cư trú.
3. Phạt tiền từ 2.000.000 đồng đến 4.000.000 đồng đối với một trong những hành vi sau đây:
a) Khai man, giả mạo hồ sơ, giấy tờ để được đăng ký thường trú, tạm trú, cấp sổ hộ khẩu, sổ tạm trú;
b) Làm giả sổ hộ khẩu, sổ tạm trú hoặc giả mạo điều kiện để được đăng ký thường trú;
c) Sử dụng sổ hộ khẩu, sổ tạm trú giả;
d) Cho người khác đăng ký cư trú vào chỗ ở của mình để vụ lợi hoặc trong thực tế người đăng ký cư trú không sinh sống tại chỗ ở đó;
đ) Cá nhân, chủ hộ gia đình cho người khác nhập hộ khẩu vào cùng một chỗ ở của mình nhưng không bảo đảm diện tích tối thiểu trên đầu người theo quy định;
e) Ký hợp đồng lao động không xác định thời hạn với người lao động không thuộc doanh nghiệp của mình để nhập hộ khẩu;
g) Sử dụng hợp đồng lao động trái với quy định của pháp luật để nhập hộ khẩu;
h) Không khai báo tạm trú cho người nước ngoài thuê nhà để ở.
4. Hình thức xử phạt bổ sung:
Tịch thu tang vật, phương tiện vi phạm hành chính đối với hành vi quy định tại Điểm a Khoản 2; Điểm a, b, c Khoản 3 Điều này.
5. Biện pháp khắc phục hậu quả:
a) Buộc thu hồi sổ hộ khẩu, sổ tạm trú, giấy tờ khác liên quan đến cư trú đối với hành vi quy định tại Điểm a Khoản 2; Điểm a Khoản 3 Điều này;
b) Buộc hủy bỏ thông tin, tài liệu sai sự thật đối với hành vi quy định tại Điểm b Khoản 2 Điều này;
c) Buộc nộp lại số lợi bất hợp pháp có được do thực hiện hành vi vi phạm hành chính quy định tại Điểm d Khoản 3 Điều này;
d) Buộc hủy bỏ hợp đồng lao động trái quy định của pháp luật để nhập hộ khẩu quy định tại Điểm e, g Khoản 3 Điều này.

Reason: The second answer does not answer the question "bao nhiêu tiền"; therefore, the retrieval system ignores it. Moreover; the second retrieved article seems to be relevant to the question (red line) but is not labeled to be the answer.

5 Conclusion and future works

5.1 Conclusion

In this project, we work on designing a two-stage retrieval system for the ZALO AI Challenge Legal Text Retrieval which include a dual encoder retriever and a cross-encoder classifier.

The sub-retriever and classifier is fine-tuned on PhoBERT[8] architecture. Our final retriever produces competitive results on the test set.

5.2 Future works

With the help of faiss similarity search[4], dual encoder has become the ideal article-filter in this challenge. However, there still some remaining matters like the lacking training data and resources as well as the limited ability of PhoBERT to handle with long articles. In the futures, we will continue work on these problems.

On the other hand, using cross-encoder phoBERT classifier leads to slow running time despite producing good results. Finding another methods to evaluated filtered articles is still a crucial task to make the system more efficient and real-life applicative.

Refercences

[1] Jacob Devlin andothers. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. inProceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers): Minneapolis, Minnesota: Association for Computational Linguistics, june 2019, pages 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[2] Luyu Gao and Jamie Callan. Condenser: a Pre-training Architecture for Dense Retrieval. 2021. arXiv: 2104.08253 [cs.CL].

[3] Luyu Gao andothers. Scaling Deep Contrastive Learning Batch Size under Memory Limited Setup. 2021. arXiv: 2101.06983 [cs.LG].

[4] Jeff Johnson, Matthijs Douze and Hervé Jégou. “Billion-scale similarity search with GPUs”. inIEEE Transactions on Big Data: 7.3 (2019), pages 535–547.

[5] Vladimir Karpukhin andothers. “Dense Passage Retrieval for Open-Domain Question Answering”. inProceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): Online: Association for Computational Linguistics, november 2020, pages 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550. URL: https://www.aclweb.org/anthology/2020.emnlp-main.550.

[6] Yinhan Liu andothers. RoBERTa: A Robustly Optimized BERT Pretraining Approach. cite arxiv:1907.11692. 2019. URL: http://arxiv.org/abs/1907.11692.

[7] Yuanhua Lv and ChengXiang Zhai. “Lower-Bounding Term Frequency Normalization”. inProceedings of the 20th ACM International Conference on Information and Knowledge Management: CIKM ’11. Glasgow, Scotland, UK: Association for Computing Machinery, 2011, pages 7–16. ISBN: 9781450307178. DOI: 10.1145/2063576.2063584. URL: https://doi.org/10.1145/2063576.2063584.

[8] Dat Quoc Nguyen and Anh Tuan Nguyen. “PhoBERT: Pre-trained language models for Vietnamese”. inFindings of the Association for Computational Linguistics: EMNLP 2020: 2020, pages 1037–1042.

[9] Yingqi Qu andothers. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. 2021. arXiv: 2010.08191 [cs.CL].

[10] Andrew Trotman, Antti Puurula and Blake Burgess. “Improvements to BM25 and Language Models Examined”. inAustralasian Document Computing Symposium: 2014.

[11] Thanh Vu andothers. “VnCoreNLP: A Vietnamese Natural Language Processing Toolkit”. inProceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations: New Orleans, Louisiana: Association for Computational Linguistics, june 2018, pages 56–60. DOI: 10.18653/v1/N18-5012. URL: https://aclanthology.org/N18-5012.