# BACHELOR THESIS

## The proposed Dual Encoder model for Open-domain question answering system: Case study in Vietnamese COVID-19 topic

*Author* Le Vu Loi - 20173240
*Supervisor* Assoc. Prof. Pham Van Hai

June 27, 2021

# Outline

# Open-domain question answering

**Open-domain question answering is at medium level of difficulty among various NLP tasks.**

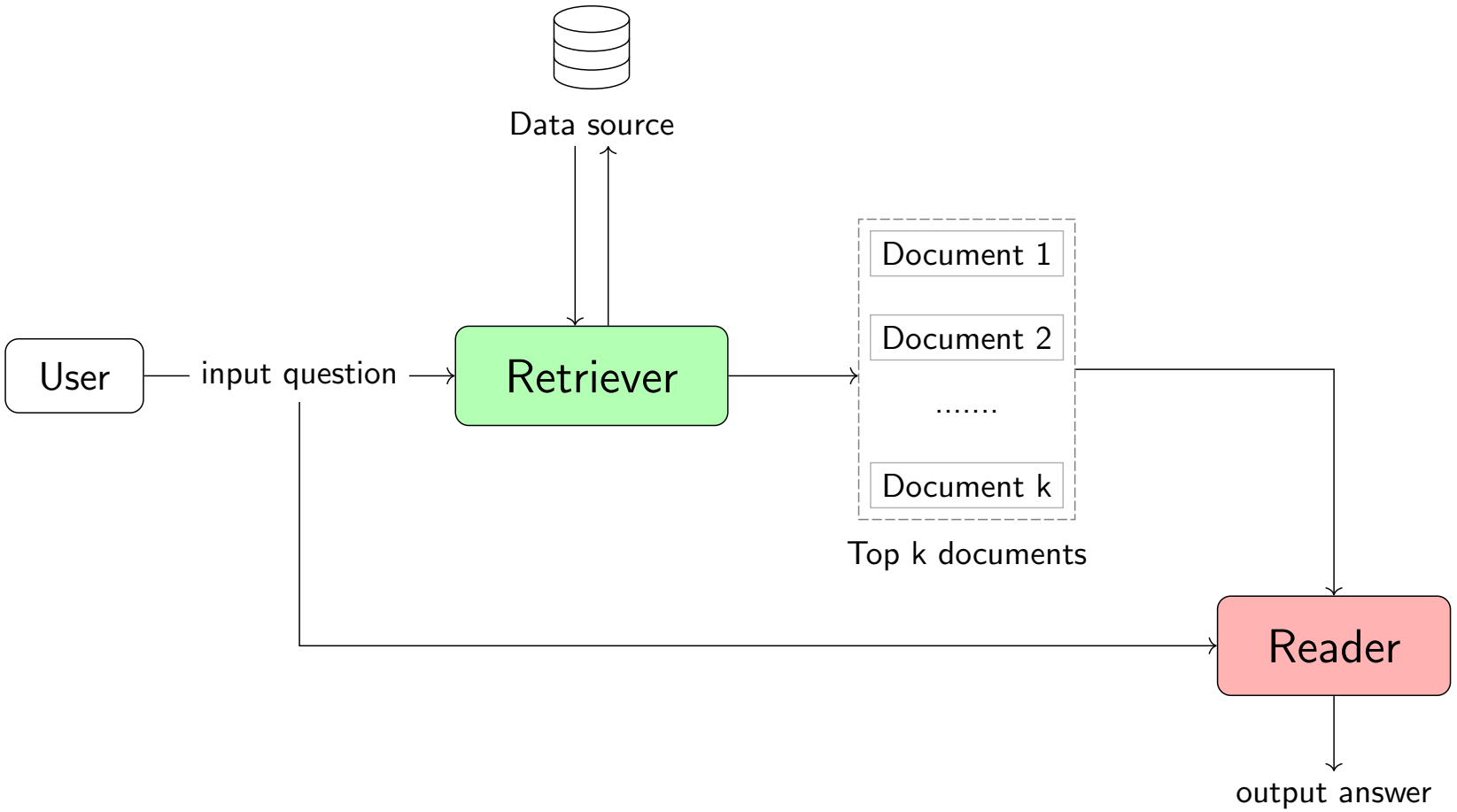
- Word embedding
- Sentence embedding
- Language Modeling

  .....
- Question Answering
- **Open-domain question answering**

  .....
- Text Summarization
- Dialogue Management

# Open-domain question answering

- Combination of **Information Retrieval** and **Machine Reading Comprehension**



# Problem formulation

- **Input**
  - A question in human natural language.

    *E.g.* Who is the founder of Google?
- **Output**
  - A list of answers for the input question

    *E.g.* [Larry Page, Sergey Brin]
- **Constraints**
  - The system answers only factoid question.
  - Factoid question means that the question is about a fact and often can be answered by a short phrase. Yes/no question, multiple-choice question and reasoning question are not factoid.

    - *Factoid question*: What is the capital of Vietnam?
    - *Reasoning question*: If 3 cats can catch 3 mice in 3 minutes, how many mices can 6 cats catch in 6 minutes?

# Related works


[1] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051, 2017.

[2] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906, 2020.


# Reading Wikipedia to answer open-domain questions

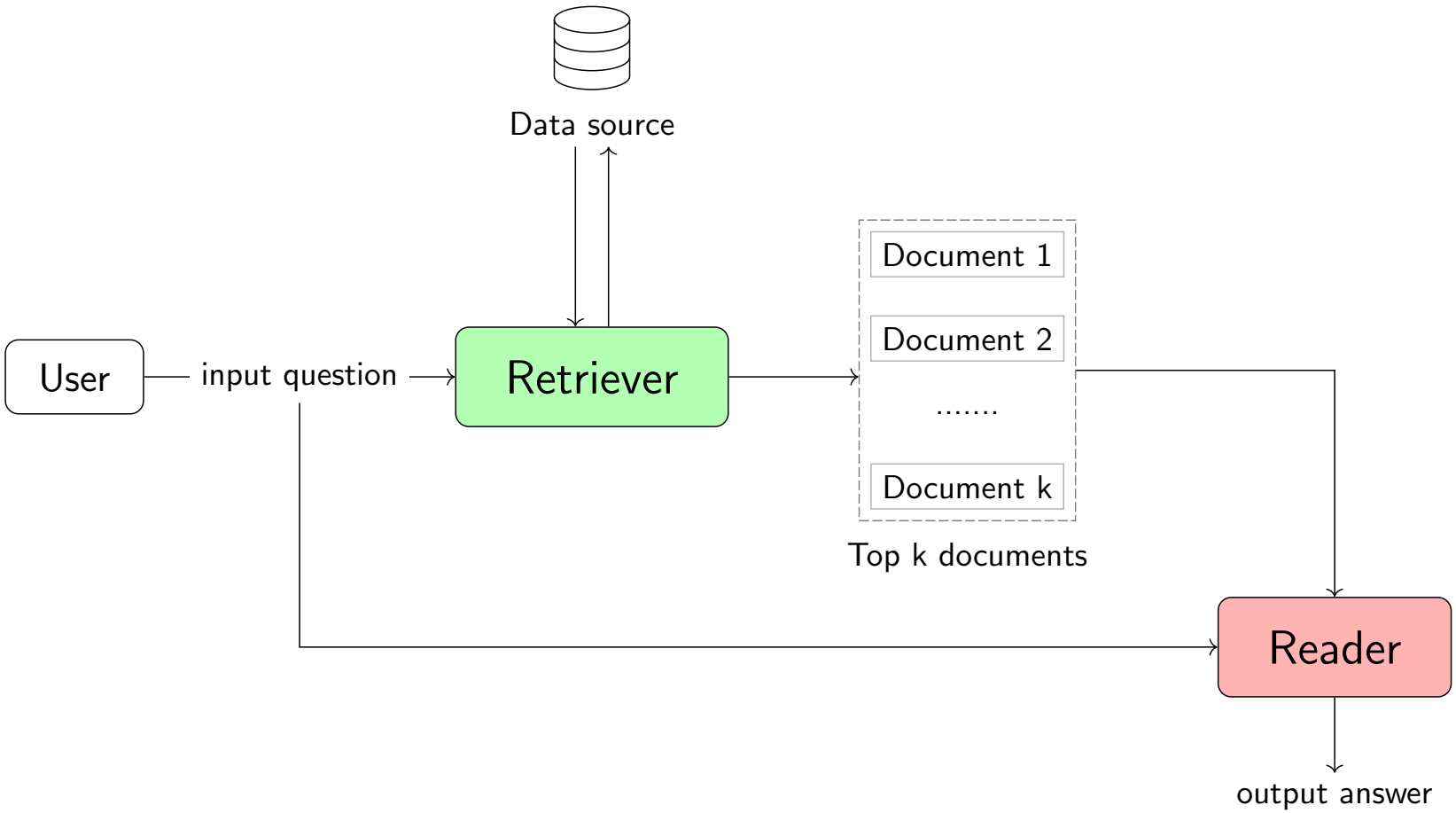
- Danqi Chen et. al [1] proposed to solve open-domain question answering problem by reading Wikipedia to find answers.
- Consist of a **document retriever** based on bigram hashing and TF-IDF matching and **machine reader** based on multi-layer Recurrent neural network.
- This work promoted a large number of subsequent publications on Open-domain question answering problem.

# Dense passage retrieval

- Karpukhin et. al [2] proposed to use a dense retriever which based on Dual encoder architecture for retrieving documents in open-domain question answering system.
- Dual encoder consists of a question encoder and a context encoder, which is used to encode question and document respectively into (a) vector space(s).
- Similarity between a document and a question is computed by taking dot product of the encoded question and the encoded document.
- Relevant documents to the input question are retrieved using this similarity.

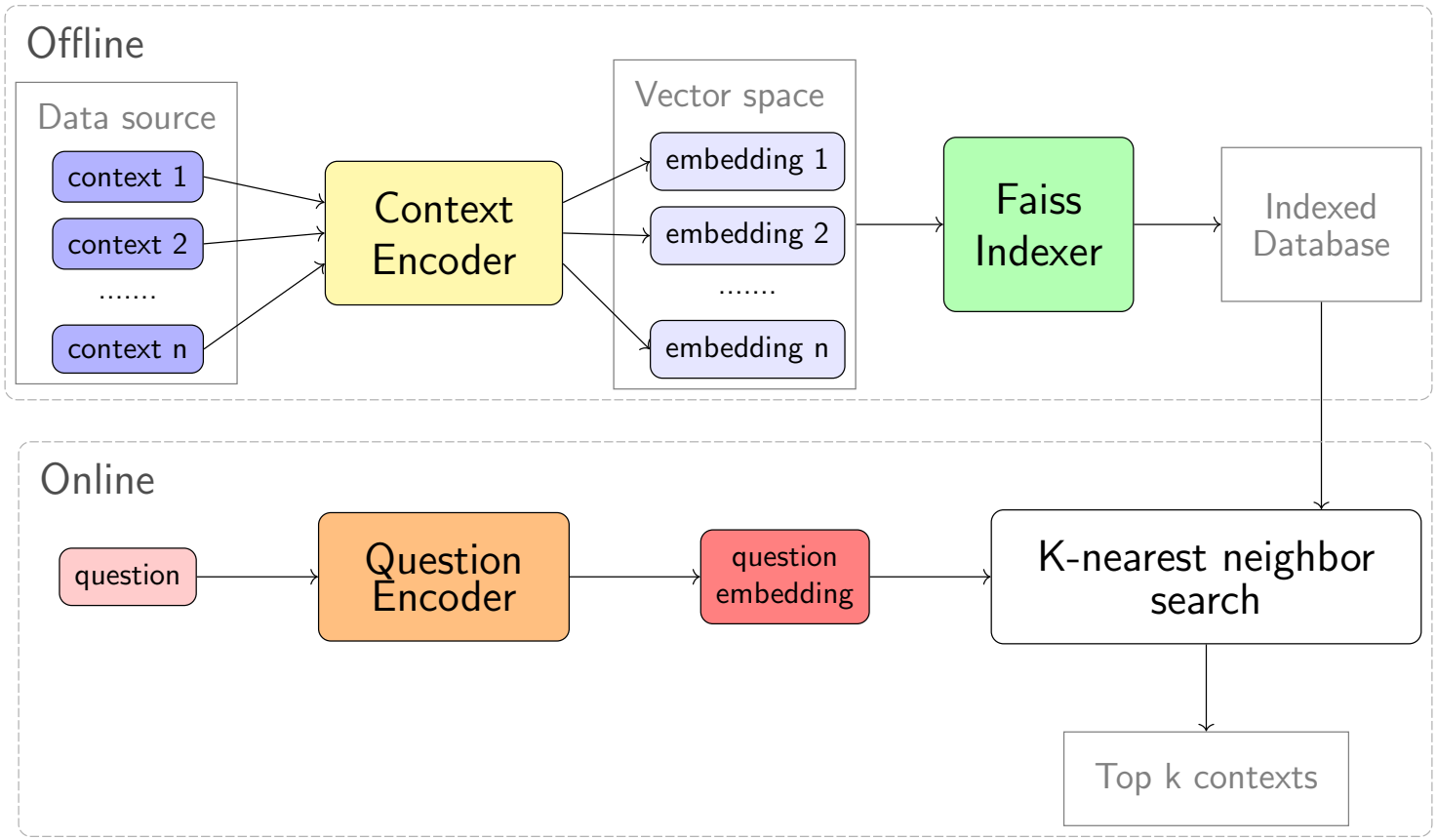
# System pipeline

- Open-domain question answering = Retriever + Reader



# Dense retriever: Dual encoder architecture

- Dense retriever is based on Dual encoder architecture.




Workflow of a dense retriever

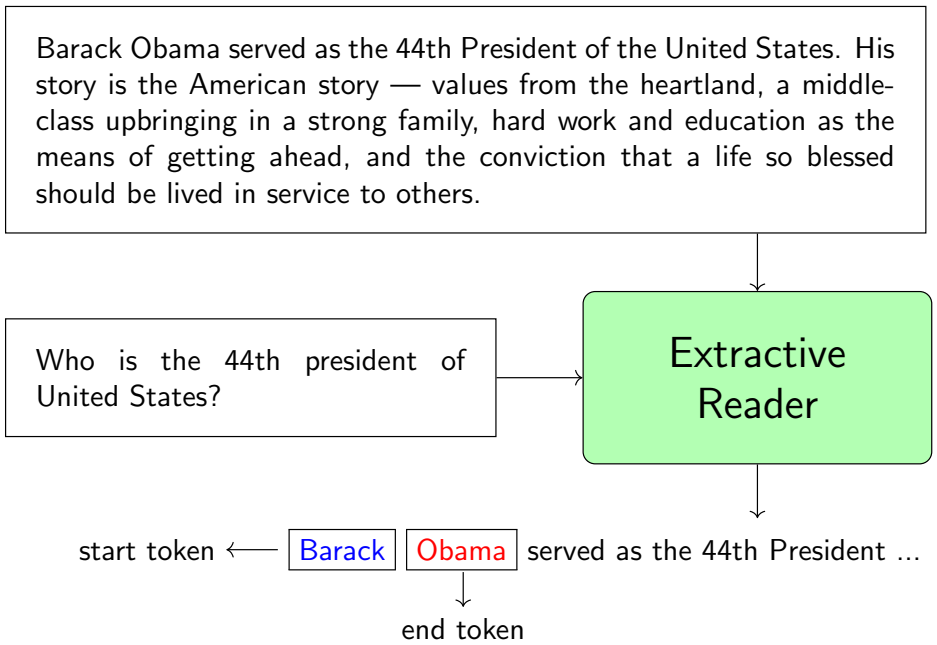
# Training dense retriever

- Jointly train question encoder and context encoder.
- Training data: a training sample consists of:
  - $q$: input question.
  - $p^+$: positive context, which is the document that contains the answers.
  - $\left\{p_j^-\right\}_{j=1}^m$: $m$ negative contexts, which are documents that do not contain the answers.
- Loss function (per one training sample): negative log-likelihood

$$\mathcal{L} = -\log\left\{\frac{\exp\left[\text{sim}\left(q, p^+\right)\right]}{\exp\left[\text{sim}\left(q, p^+\right)\right] + \sum_{j=1}^{m}\exp\left[\text{sim}\left(q, p_j^-\right)\right]}\right\} \quad (1)$$

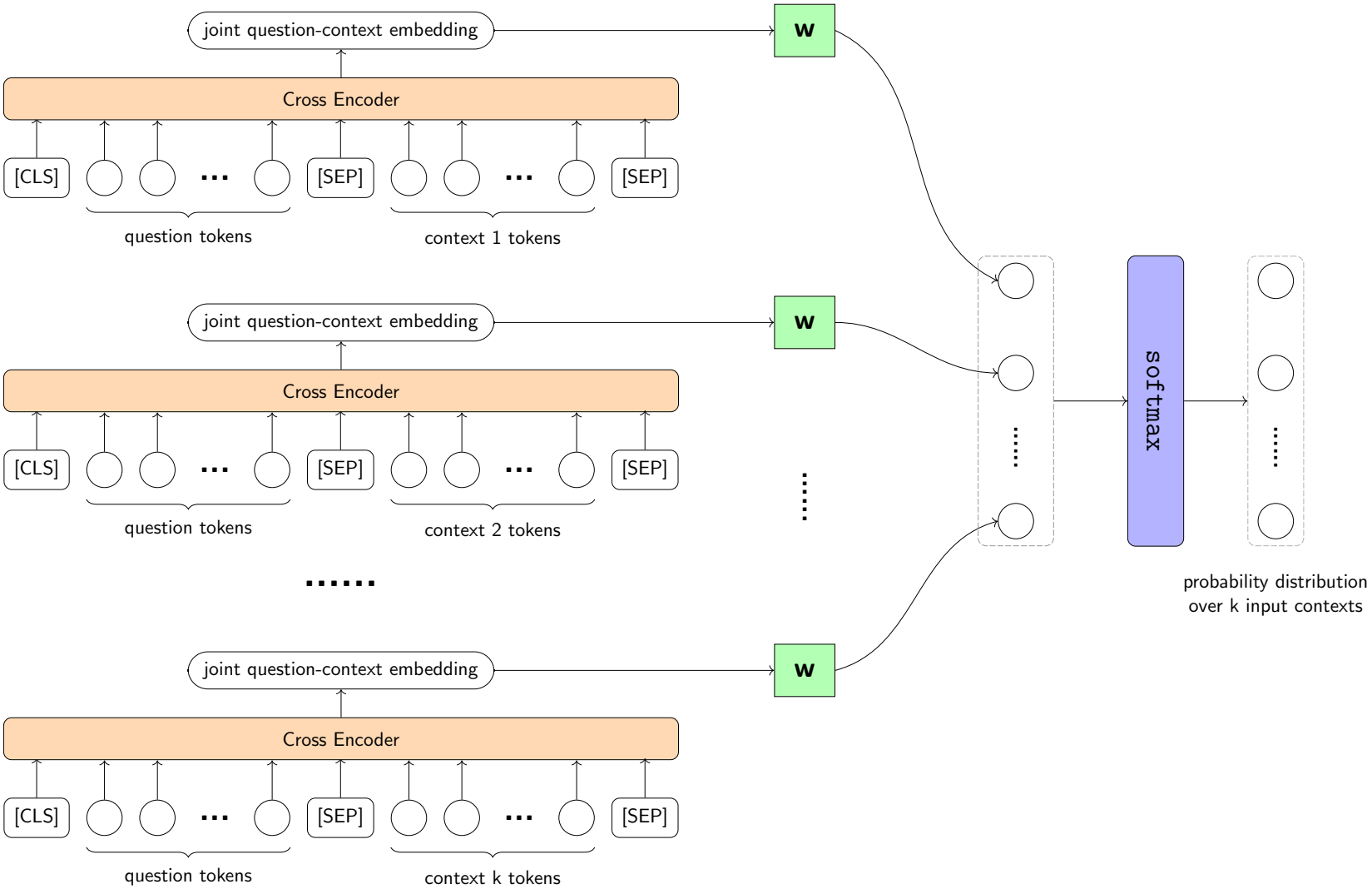
# Extractive reader: Cross encoder architecture

- Extractive reader's task is to predict the start and end position of answer in the documents returned by dense retriever.



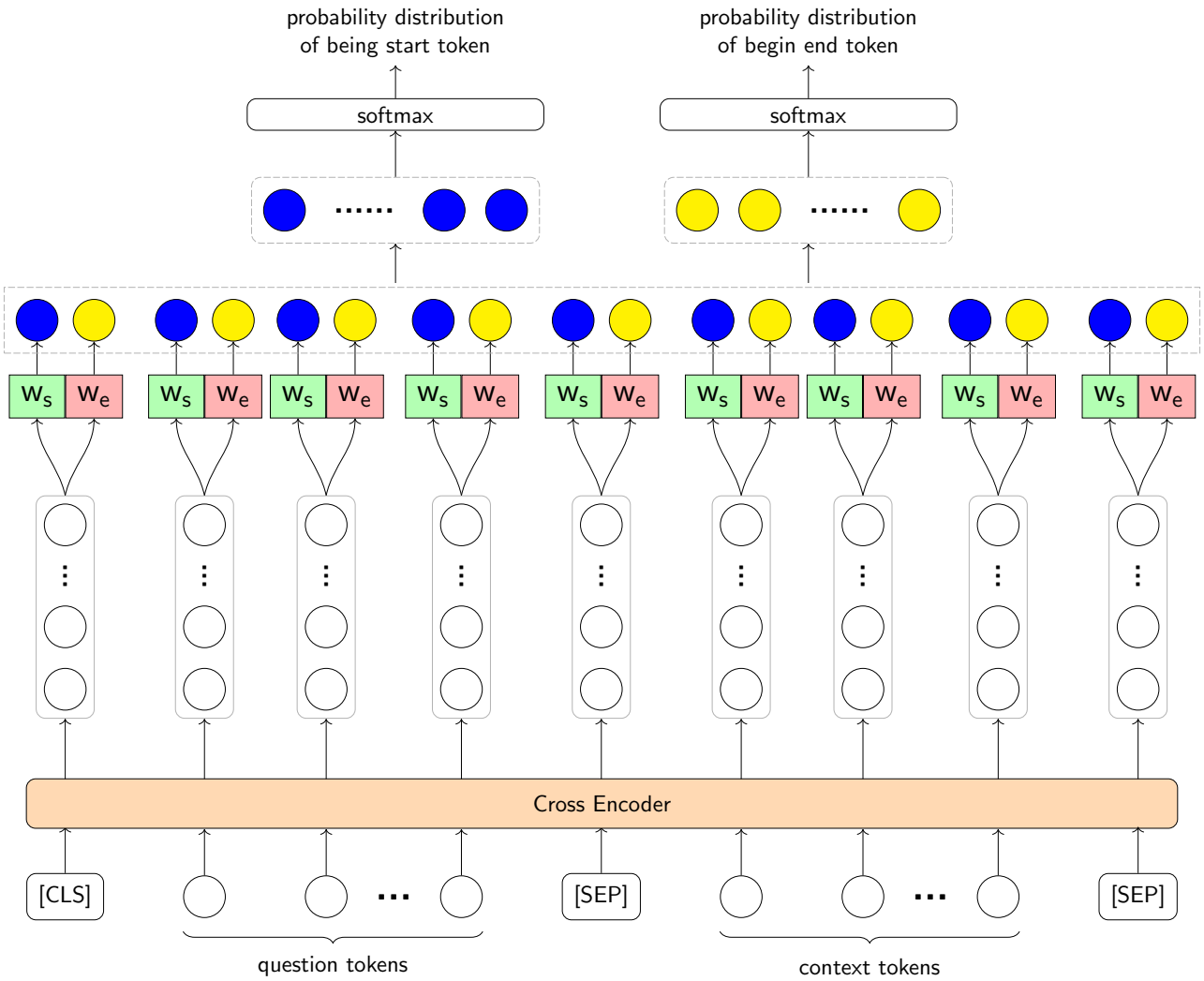

- Extractive reader consists of 2 components, in which each component follows a cross encoder architecture:
  - Re-ranker: re-rank documents returned by dense retriever.
  - Single-document reader: read one document to extract answers.

# Re-ranker




Model architecture

# Single-document reader




Model architecture

# Existing approach: in-batch loss

- Karpukhin et. al [2] proposed to use in-batch strategy and hard negative contexts to train dense retriever using negative log-likelihood loss defined in equation (1). To be specific:
  - In-batch strategy: training samples in a training batch use others' positive context as their negative contexts
    $\rightarrow$ significantly reduce the number of documents needed to be fed into context encoder.
  - Hard negative contexts: normal negative contexts are non-relevant to the input question. Hard negative contexts are relevant to the input question but do not contain required information to answer that question.

    $\rightarrow$ challenge the model to learn better.

# Proposed method: stratified loss for training dual encoder

- Idea: additional loss for learning difference between hard negative and normal negative contexts.
- Stratified loss
  - Assumptions: a batch of $b$ training samples $\mathcal{D}$, where the $i$-th training sample $\mathcal{D}_i$ consists of:
    - $q_i$: input question.
    - $p_i^+$: positive context.
    - $\left\{p_{i,j}^-\right\}_{j=1}^m$: $m$ hard negative contexts.
  - Loss formula

$$\begin{aligned}\mathcal{L} = & -\log\left\{\frac{\exp\left[\text{sim}\left(q_i, p_i^+\right)\right]}{\exp\left[\text{sim}\left(q_i, p_i^+\right)\right] + \sum_{j=1}^{m}\exp\left[\text{sim}\left(q_i, p_{i,j}^-\right)\right]}\right\} \\ & -\sum_{j=1}^{m}\log\left\{\frac{\exp\left[\text{sim}\left(q_i, p_{i,j}^-\right)\right]}{\exp\left[\text{sim}\left(q_i, p_{i,j}^-\right)\right] + \sum_{k\in\{1,2,\ldots,b\}\setminus\{i\}}\exp\left[\text{sim}\left(q_i, p_k^+\right)\right]}\right\}\end{aligned} \tag{2}$$

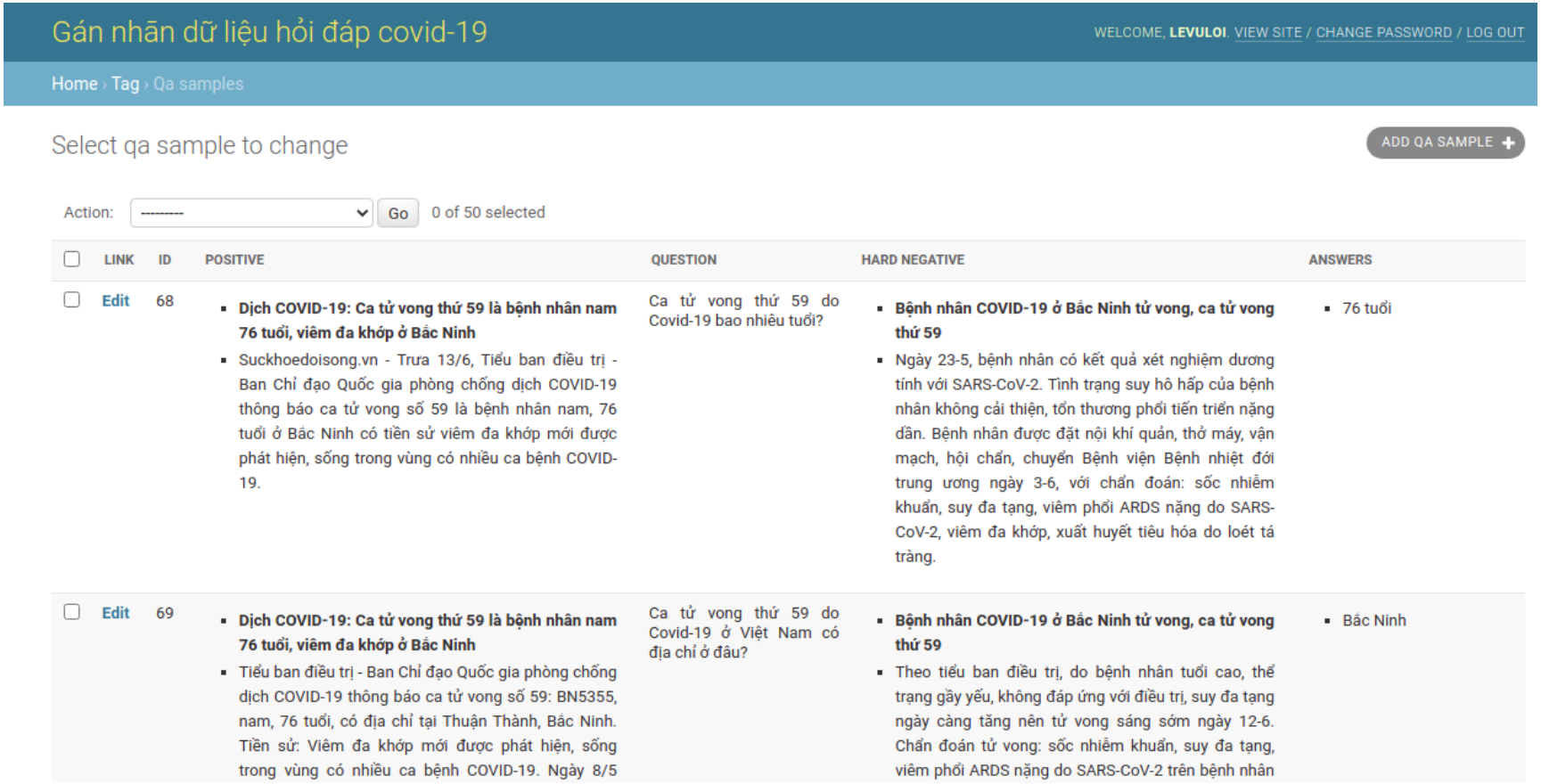
# Case study on Vietnamese COVID-19 topic

- Building an open-domain question answering for Vietnamese COVID-19 topic requires:
  - Building a context source for COVID-19 topic, which contains all documents that the system searches during answering a question about COVID-19 topic.
  - Annotate data for training dense retriever and extractive reader (re-ranker and single-document reader).

# Data crawling for COVID-19 data

- Context source: 168,388 contexts/documents about medial topic, mainly crawled from `https://suckhoedoisong.vn/`
- Training data: 995 training samples, in which each sample consists of:
  - Input question
  - One positive context
  - One hard negative context
  - List of answers

# Data annotating



Gán nhãn dữ liệu hỏi đáp covid-19

WELCOME, **LEVULOI**. VIEW SITE / CHANGE PASSWORD / LOG OUT

Home › Tag › Qa samples

Select qa sample to change

ADD QA SAMPLE +

Action: -------- Go 0 of 50 selected

| | LINK | ID | POSITIVE | QUESTION | HARD NEGATIVE | ANSWERS |
|---|---|---|---|---|---|---|
| ☐ | Edit | 68 | **Dịch COVID-19: Ca tử vong thứ 59 là bệnh nhân nam 76 tuổi, viêm đa khớp ở Bắc Ninh**<br>Suckhoedoisong.vn - Trưa 13/6, Tiểu ban điều trị - Ban Chỉ đạo Quốc gia phòng chống dịch COVID-19 thông báo ca tử vong số 59 là bệnh nhân nam, 76 tuổi ở Bắc Ninh có tiền sử viêm đa khớp mới được phát hiện, sống trong vùng có nhiều ca bệnh COVID-19. | Ca tử vong thứ 59 do Covid-19 bao nhiêu tuổi? | **Bệnh nhân COVID-19 ở Bắc Ninh tử vong, ca tử vong thứ 59**<br>Ngày 23-5, bệnh nhân có kết quả xét nghiệm dương tính với SARS-CoV-2. Tình trạng suy hô hấp của bệnh nhân không cải thiện, tổn thương phổi tiến triển nặng dần. Bệnh nhân được đặt nội khí quản, thở máy, vận mạch, hội chẩn, chuyển Bệnh viện Bệnh nhiệt đới trung ương ngày 3-6, với chẩn đoán: sốc nhiễm khuẩn, suy đa tạng, viêm phổi ARDS nặng do SARS-CoV-2, viêm đa khớp, xuất huyết tiêu hóa do loét tá tràng. | 76 tuổi |
| ☐ | Edit | 69 | **Dịch COVID-19: Ca tử vong thứ 59 là bệnh nhân nam 76 tuổi, viêm đa khớp ở Bắc Ninh**<br>Tiểu ban điều trị - Ban Chỉ đạo Quốc gia phòng chống dịch COVID-19 thông báo ca tử vong số 59: BN5355, nam, 76 tuổi, có địa chỉ tại Thuận Thành, Bắc Ninh. Tiền sử: Viêm đa khớp mới được phát hiện, sống trong vùng có nhiều ca bệnh COVID-19. Ngày 8/5 | Ca tử vong thứ 59 do Covid-19 ở Việt Nam có địa chỉ ở đâu? | **Bệnh nhân COVID-19 ở Bắc Ninh tử vong, ca tử vong thứ 59**<br>Theo tiểu ban điều trị, do bệnh nhân tuổi cao, thể trạng gầy yếu, không đáp ứng với điều trị, suy đa tạng ngày càng tăng nên tử vong sáng sớm ngày 12-6. Chẩn đoán tử vong: sốc nhiễm khuẩn, suy đa tạng, viêm phổi ARDS nặng do SARS-CoV-2 trên bệnh nhân | Bắc Ninh |

# Pretrained model

- Vietnamese open-domain question answering for COVID-19 topic was trained from the pretrained language model `NlpHUST/vibert4news-base-cased` pulling from `huggingface`
- Another well-known pretrained language model for Vietnamese is `PhoBERT` of VinAI, which can also be found on `huggingface`

# Datasets

- Google Natural Question: preprocessed data taken from [2]
  - 58,880 training samples
  - 8,757 development samples
  - 3,610 test samples
  - Context source contains 21,015,324 contexts
  - To rapidly produce experiments, the context source is reduced to 700,000 contexts and 450 additional contexts are considered to cover all input questions in the test set.
- Vietnamese COVID-19 datatset
  - 995 training samples
  - Context source contains 168,388 contexts

# Metrics

- Top-$k$ hit scores
  - Measure retriever's accuracy
  - Top-$k$ hit is reached if at least one of $k$ contexts returned by the retriever contains answer(s) for input question.
- Exact match
  - Measure reader's accuracy
  - Measure end-to-end system's accuracy
  - An exact match hit is reached if answer(s) produced by the open-domain question answering system matches exactly the ground truth answer(s)

# System settings

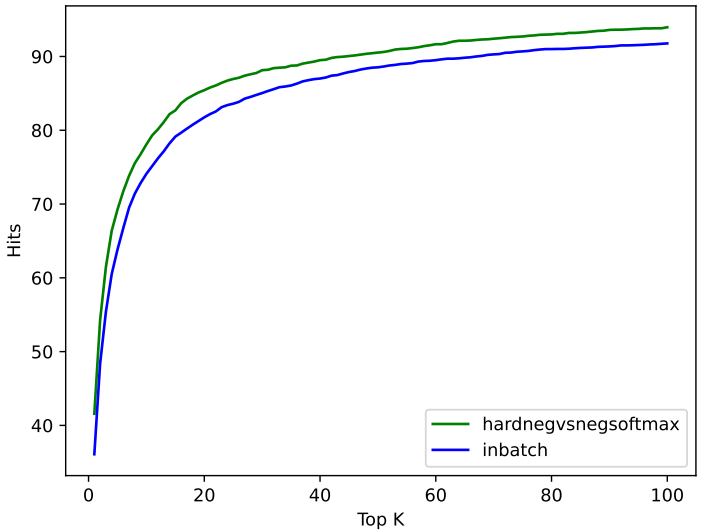
- Using Google Cloud Platform
- Training and inference on Cloud TPUs
- Process data on VM Compute Engine

Table: Hardware configurations

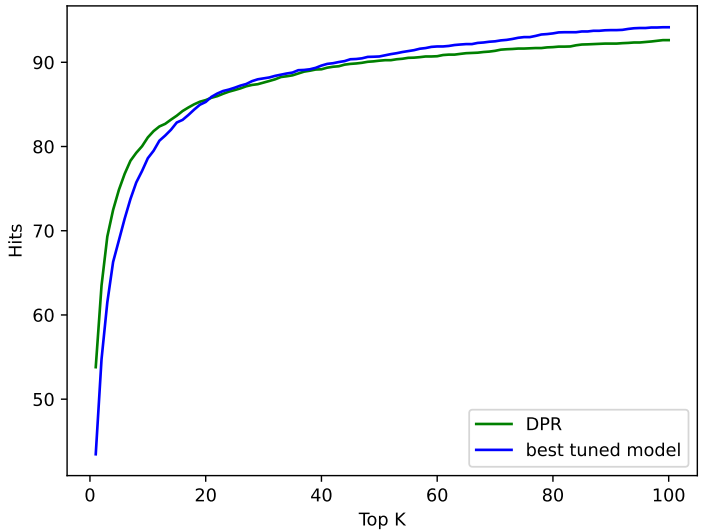
| Cloud TPUs | VM Compute Engine |
|---|---|
| TPU v3-8 on-demand:<br>• TPU version 3<br>• 8 TPU cores<br>• 16GiB memory / TPU core | • OS: Ubuntu 20.04<br>• Disk: 30GB<br>• RAM: 16GB<br>• nCPUs: 4 |

# Results on dense retriever

- Experimental results on dense retriever are conducted using Google Natural Question dataset.
- The proposed method was compared to the baseline in [2]




Comparasion results with baseline model implemented




Comparasion results with baseline model taken from released checkpoint

# Results on extractive reader

- Experimental results on extractive reader are conducted using Google Natural Question.
- Exact match score: 56.6%

# Question Answering about COVID-19

**Vietnamese Open-domain question answering for COVID-19 topic**

**Question**

Loài động vật nào là nguồn gốc của covid-19 | Go

**Long cốt trị mất ngủ, ra mồ hôi trộm**

Long cốt là xương đã hoá thạch của động vật cổ đại thuộc loài khủng long: Tê giác ngựa 3 ngón chân Rhinoceros sinensis Owen.; Rhinoceros indet; Loài hươu: Cervidae indet; loài trâu: Bovidae indet...; Long xỉ (Dens Draconis) là một loại long cốt, có cùng chung thành phần hóa học và công dụng.

**Các loài động vật đứng đằng sau dịch COVID-19**

Coronavirus hiếm khi nhảy trực tiếp từ dơi sang người, vì thế, các nhà vi trùng học đang tìm kiếm cách để giải mã cho một câu đố hóc búa: Loài vật nào đóng vai trò vật chủ trung gian giữa dơi và con người? Cầy hương được cho là vật chủ trung gian của bệnh SARS. Lạc đà là vật chủ trung gian của MERS. Trong quá trình tìm kiếm thủ phạm gây ra COVID-19, loài rắn cũng được nhắc đến, nhưng giả thuyết này bị loại bỏ bởi hiện tại không có bất kỳ bằng chứng nào về Coronavirus lây nhiễm từ động vật ngoài thú có vú và chim chóc với rắn.

**Truy tìm nguồn gốc SARS-COV-2 nhờ tái tạo lại lịch sử tiến hóa**

Tiếp theo, họ đã xây dựng lại lịch sử phát sinh gien cho các khu vực không tái tổ hợp và so sánh chúng với nhau để xem loại virus cụ thể nào đã tham gia vào các sự kiện tái tổ hợp trong quá khứ. Điều này tạo ra mối quan hệ tiến hóa giữa SARS-CoV-2 và các chủng gần nhất như từng được biết đến, đó là virus dơi và virus ở tê tê. Các nhà nghiên cứu phát hiện ra rằng dòng virus mà SARS-CoV-2 xuất phát từ các dòng virus có ở loài dơi, có cách đây khoảng 40-70 năm trước.

# Question Answering about COVID-19

**Vietnamese Open-domain question answering for COVID-19 topic**

**Question**

Khi nào Việt Nam có vắc xin chống covid-19 | Go

**Việt Nam không có đỉnh dịch COVID - 19!**

Còn nếu chúng ta chờ không được, đường biểu diễn bệnh cứ đi ngang hoài, chúng ta cần can thiệp bằng vắc xin. Thế giới đang cân nhắc xem xét, nếu tình hình tháng 5 tháng 6 tháng 7, dịch vẫn không xuống nổi, đến tháng 9, các nước buộc sẽ đưa vắc xin vào, không thể chờ thêm nữa. Vậy chúng ta có dự đoán khi nào hết dịch không, thưa bác sĩ? BS Trương Hữu Khanh: Thứ nhất, chúng ta làm tốt các biện pháp phòng ngừa dịch bệnh một cách quyết liệt mới có thể không còn bệnh nhân tại Việt Nam.

**Vắc xin COVID-19 made in Vietnam đầu tiên dự kiến sẽ có vào cuối tháng 9/2021**

Phó Thủ tướng nhắc lại nhận định của các chuyên gia, nhà khoa học cho rằng virus SARS-CoV-2 có thể có những biến đổi, tiếp tục tồn tại một số năm nữa. Cho đến giờ phút này nhiều khả năng các vắc xin phòng COVID-19 đều phải tiêm nhắc lại chứ không phải 1 đợt, hay 1 năm là xong. Dân số Việt Nam là 100 triệu người, vì vậy, chúng ta phải bằng các giải pháp để có vắc xin của Việt Nam, không chỉ phục vụ công tác phòng chống dịch COVID-19, mà còn chuẩn bị để ứng phó đối với những dịch bệnh có thể xảy ra trong tương lai.

**Chủ tịch Quốc hội Vương Đình Huệ: Đẩy nhanh tiến độ thử nghiệm vắc xin COVID-19 Nano Covax để có thể sớm sản xuất trong nước**

Đây là nhiệm vụ quan trọng để Việt Nam có thể chủ động phòng, chống COVID-19. Trong công tác phòng, chống dịch COVID-19 hiện nay, Chủ tịch Quốc hội nêu rõ, vắc xin là vũ khí quan trọng, mang tính quyết định sống còn đối với việc chấm dứt và chiến thắng đại dịch. "Nếu không sớm miễn dịch cộng đồng bằng việc tiêm chủng vắc xin COVID-19 sẽ rất khó để đẩy mạnh các hoạt động khác"- Chủ tịch Quốc hội nói.

# Thank you for your attention