BACHELOR THESIS

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

Le Vu Loi - 20173240

Talented class of Computer Science

Supervisor: Assoc. Prof. Pham Van Hai
Department of Information System

July 9, 2021

Outline

Open-domain question answering

Machine Reading Comprehension

Summarization

Word embedding

Dependency Parsing

Named Entity Recognition

Open-domain question answering

Part-of-speec Tagging

Machine Translation

Language Modeling

Sentiment Analysis

Dialogue Management

Sentence Embedding

Information retrieval

Keyword Extraction

Open-domain question answering

- Combination of **retriever** (Information Retrieval) and **reader** (Machine Reading Comprehension)
  - "Skim through" a large data source to find a subset of relevant documents.
  - "Swallow" each document to find the exact answer(s).

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.

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

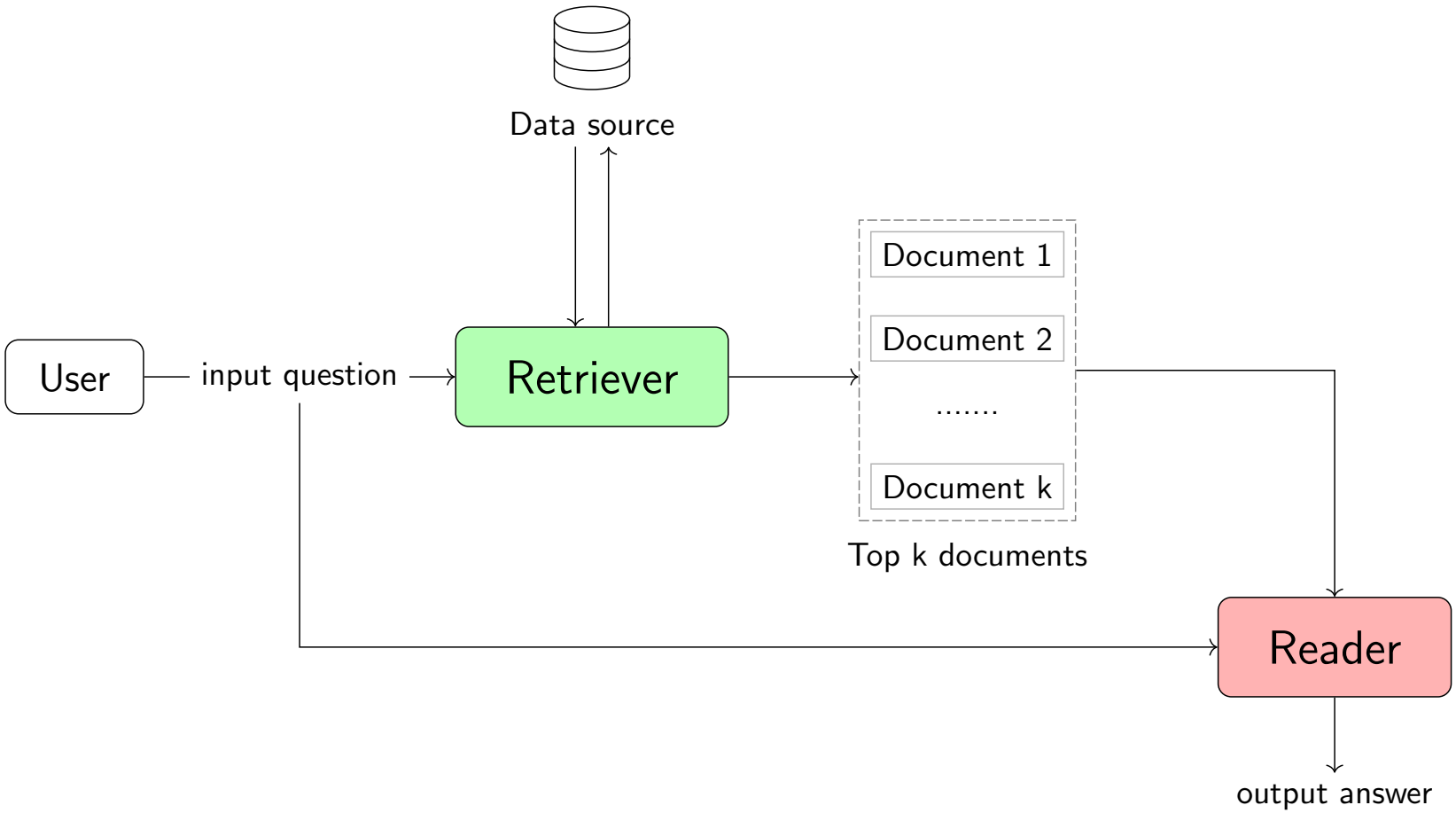
- **Retriever**: bigram hashing and TF-IDF matching
- **Reader**: Multi-layer recurrent neural network
- Potential improvements: using neural network to better capture documents' semantics

Dense passage retrieval

- **Retriever**: Dual-encoder
- **Reader**: Cross-encoder
- Successfully use neural network to solve information retrieval.
- Potential improvements: More challenging learning task for the system to gain deeper language understanding.

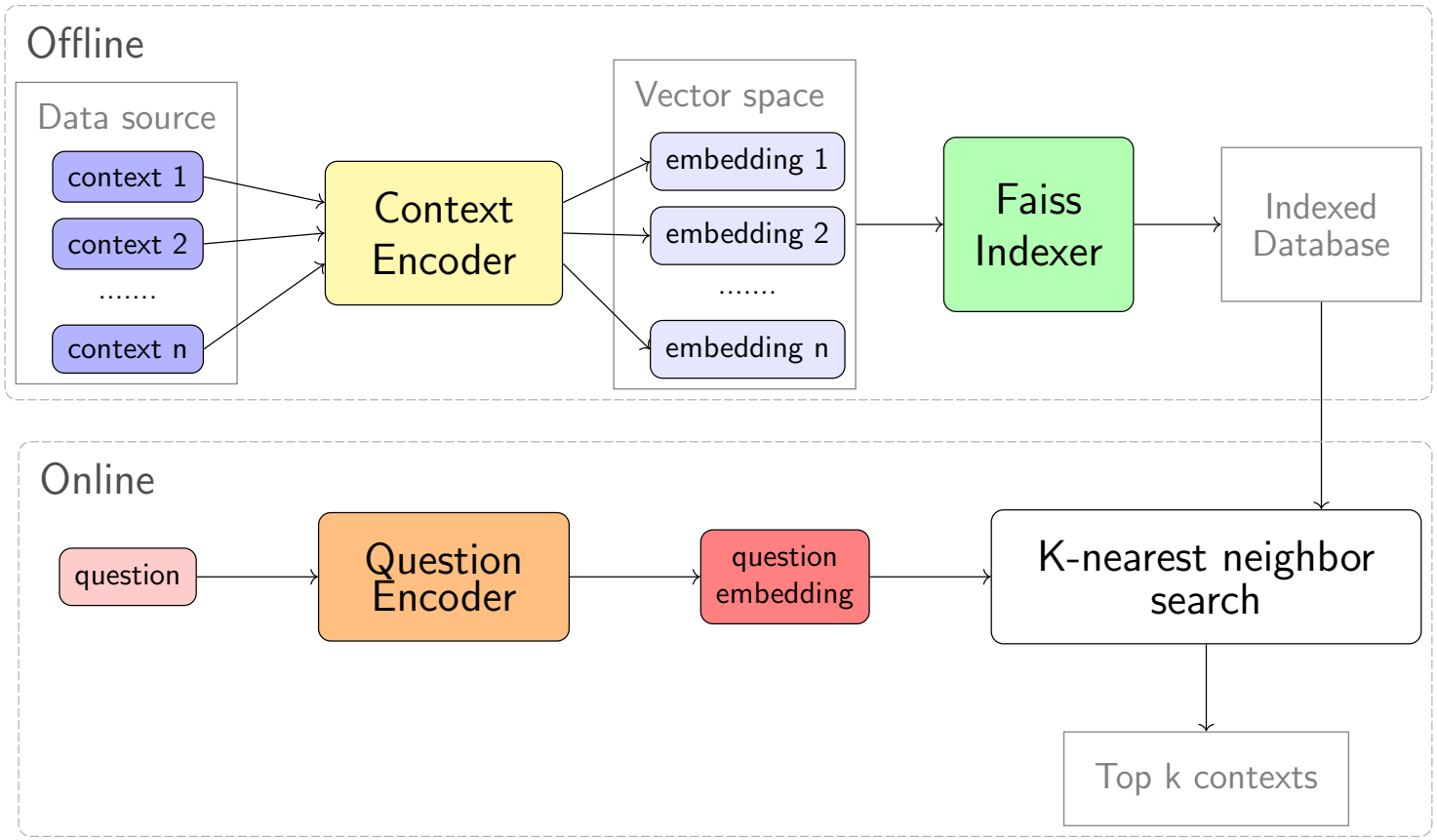
System pipeline

- Open-domain question answering = Retriever + Reader



Dense retriever: Dual encoder architecture

- Dense retriever is based on Dual encoder architecture.



Workflow of a dense retriever

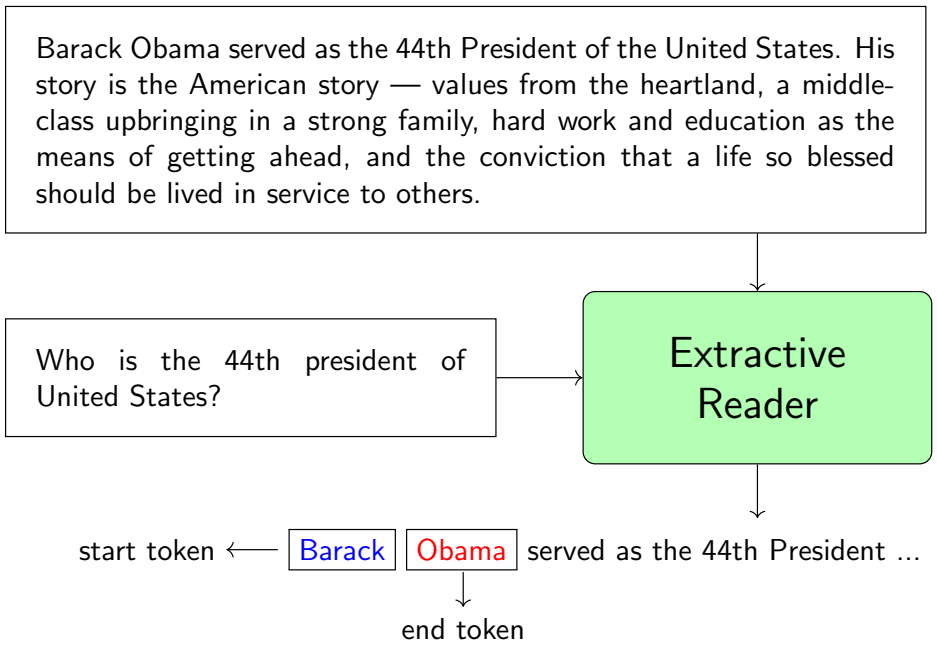
Training dense retriever

- 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

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

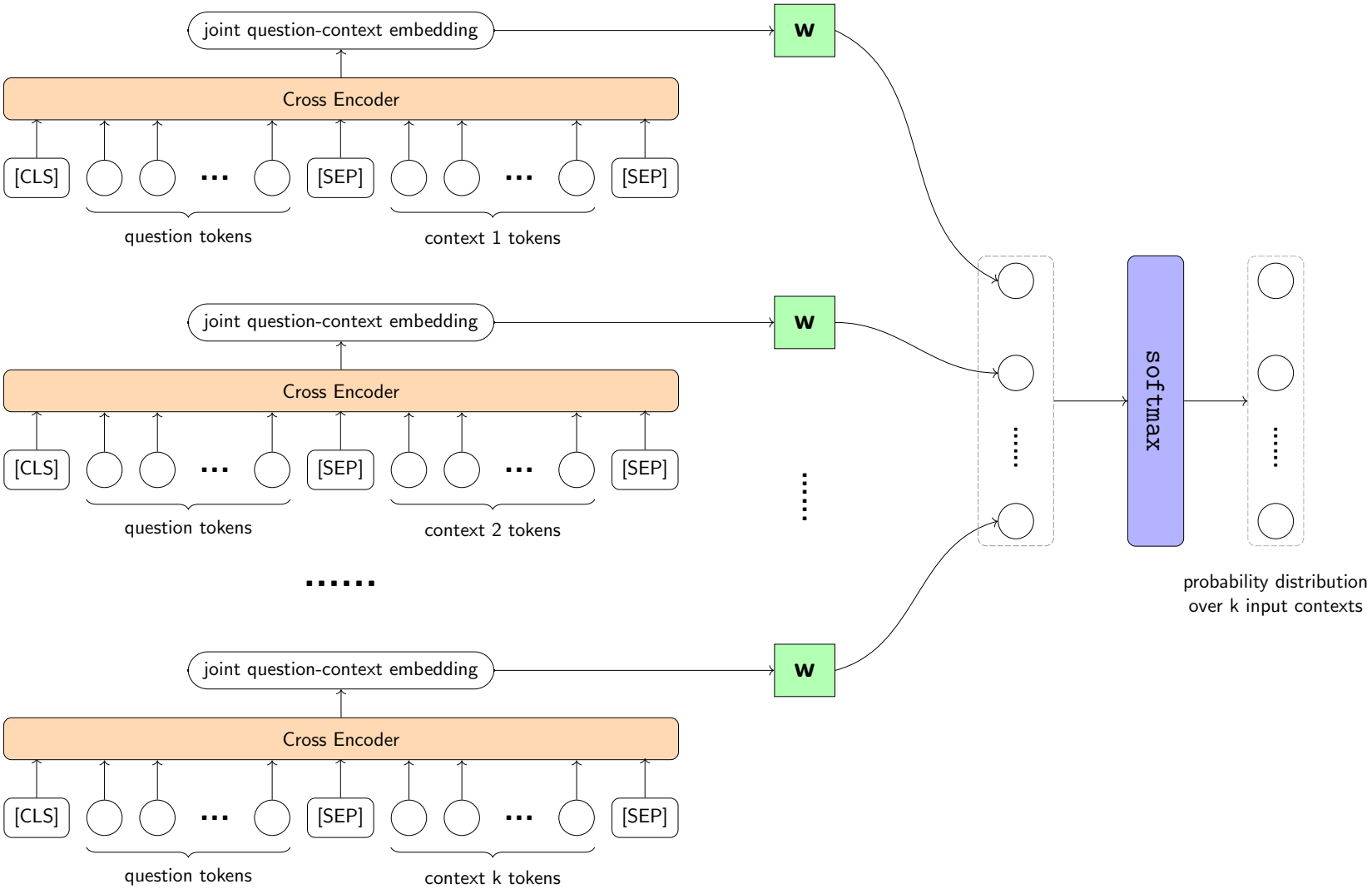
Extractive reader: Cross encoder architecture

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



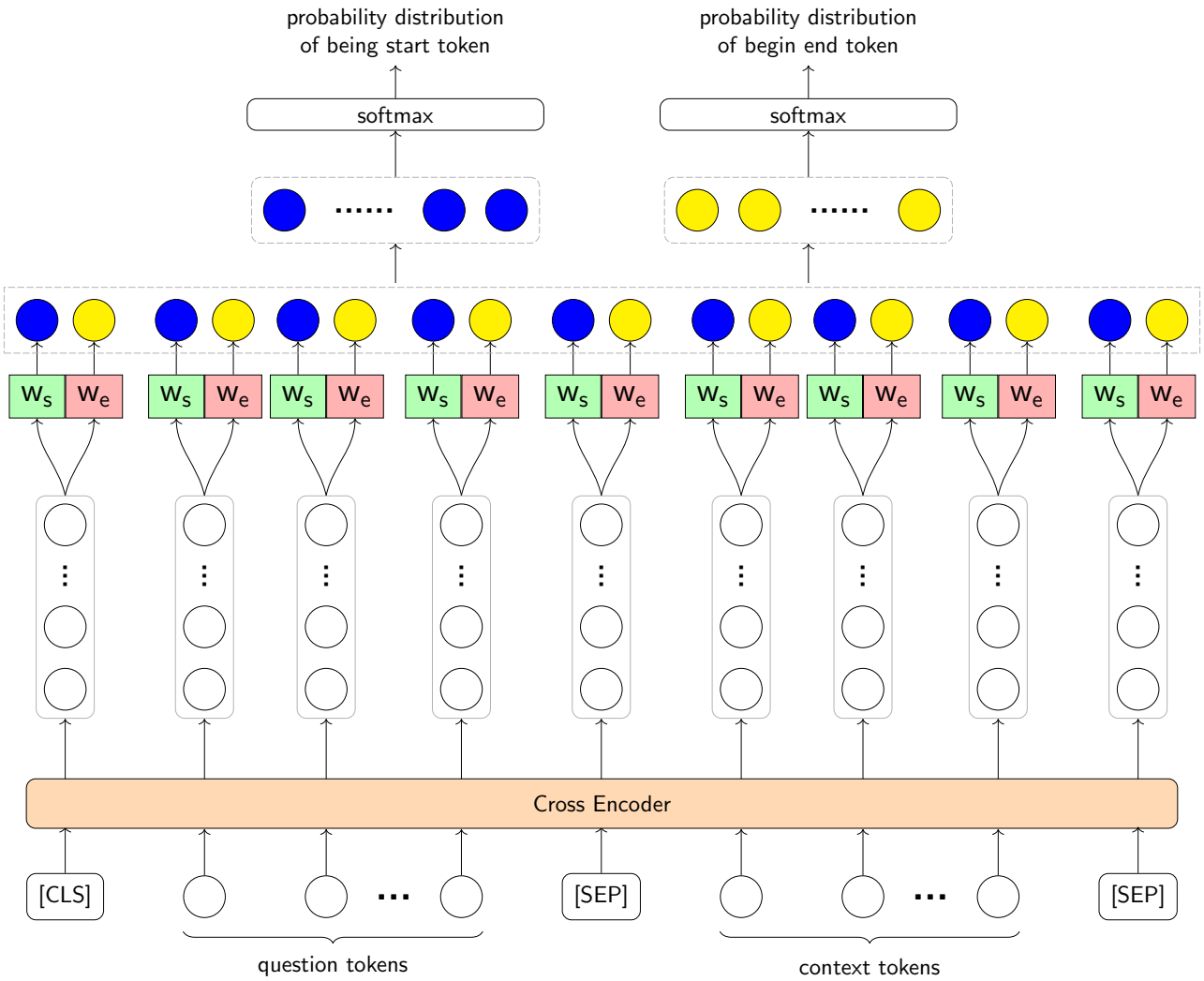
- Extractive reader consists of 2 components, in which each component follows a cross encoder architecture:
  - Re-ranker: re-rank documents returned by dense retriever.
  - Single-document reader: read one document to extract answers.

Re-ranker



Model architecture

Single-document reader



Model architecture

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

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

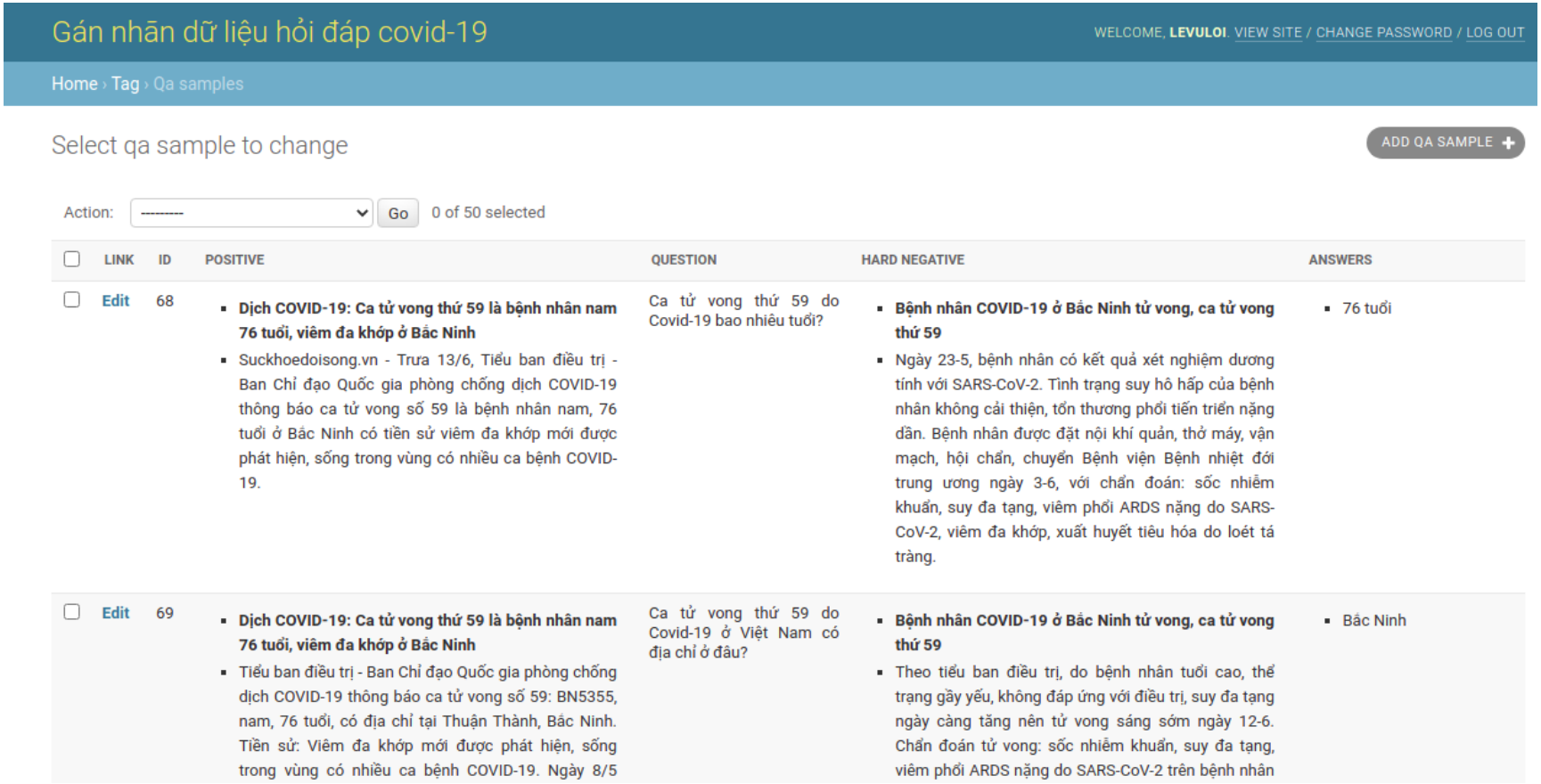
Case study on Vietnamese COVID-19 topic

- 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** ▪ 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** ▪ 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** ▪ 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** ▪ 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

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

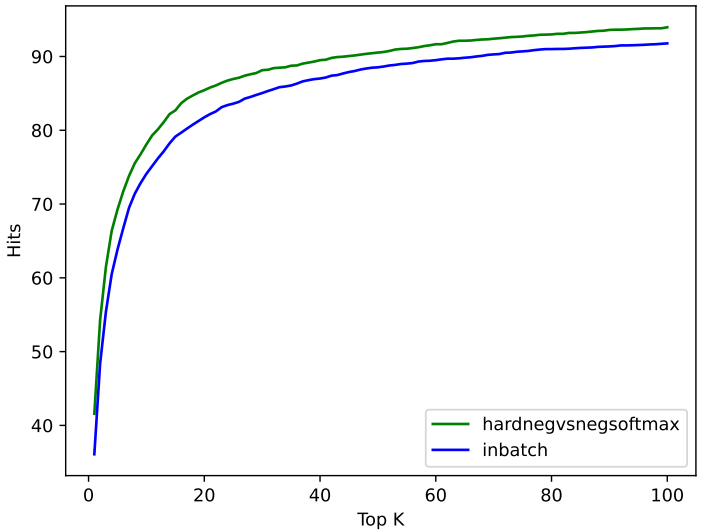
- Using Google Cloud Platform
- Training and inference on Cloud TPUs
- Process data on VM Compute Engine

Hardware configurations

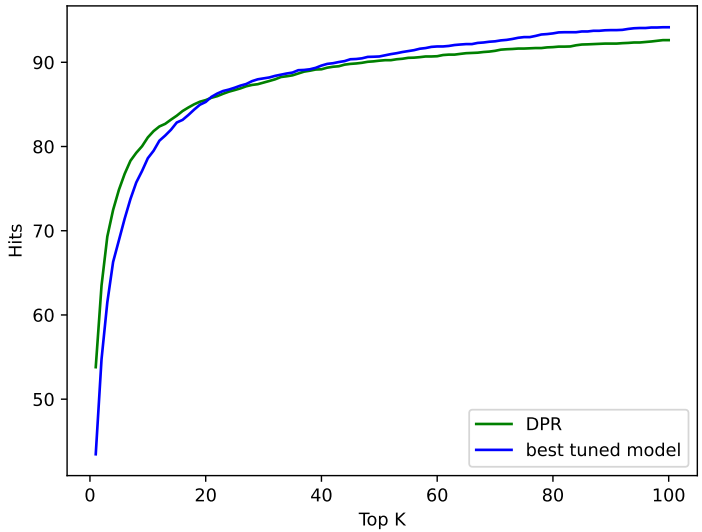
Cloud TPUs	VM Compute Engine
TPU v3-8 on-demand: • TPU version 3 • 8 TPU cores • 16GiB memory / TPU core	• OS: Ubuntu 20.04 • Disk: 30GB • RAM: 16GB • nCPUs: 4

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

Demo: 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.

Conclusion and future works

- Conclusion
  - Propose to train retriever model with stratified loss
  - Conduct a case study for open-domain question answering system in Vietnamese language for COVID-19 topic
  - Use Cloud TPUs to train large retriever model in short time
- Future works
  - Study machine reading comprehension problem to improve reader component
  - Study the relationship between open-domain question answering and automatic knowledge graph construction
  - Study dialogue management problem, which extends question answering problem by adding multi-turn conversation ability

Thank you for your attention