sBERT & BART : Models for Open Domain Long Form Question Answering

Khoa D. Vo, Dang M. Nguyen and Anh T. H. Nguyen

AI Engineer Intern, Gradients Technologies

NLP Training Lab

Mr. Hieu Pham

Sep. 20, 2023

Introduction

In the modern era, the Internet has become a popular place for anyone to search for any kind of questions, such as how to make delicious egg fried rice, or whether the chicken comes first or the egg comes first. However, the problem is the Internet is a very big place, how can you find the right information in short notice?

If you are lucky enough, and your particular query has already been asked and answered succinctly and clearly on one of the numerous question and answer websites available online (like Quora, Reddit, or Yahoo Answers): modern search engines will likely take you to that previously existing answer pretty reliably in a matter of clicks.

But in some cases, where you are the first person to raise that issue, the procedure might be a little more difficult. It is probable that you will need to gather pertinent data from a range of sources, analyse how these facts relate to your issue, and then build together a story that responds to it.

Now would it be wonderful if your computer could do all of that for you: gather the appropriate sources (for example, sentences from pertinent Wikipedia articles), synthesise the data, and draw out a simple, original summary of the important points? At least not one that can offer accurate information in its summary, such a system is not currently available. Even while current systems are excellent at locating an extractive span that responds to a factual inquiry in a given text, they still struggle with extended response generation in open-domain scenarios where a model must locate its own sources of data.

Fortunately, recent advances in the field of natural language processing have taken us a few steps closer to handling such tasks. These developments include advancements in the pre-training of information retrieval models such as sBERT as well as sequence-to-sequence models for conditional text generation (e.g. BART, T5).

In this paper, we show how we use the pre-trained sBERT: sentence Bidirectional Encoder Representations from Transformers, combined with fine-tuning BART model using LoRA: Low-Rank Adaptation of Large Language Models on the Eli5 dataset, to take in a question, fetch top 10 relevant passages from Wikipedia Snippets, then compile a multi-sentence answer based on the question and retrieved passages.

sBERT & BART : Open Domain Long Form Question Answering

The process of developing the system goes through 4 stages: data preparation, fine-tuning BART for generating answers, creating context query function with sBERT, project finalisation.

Data Preparation

In this phase, we collect data from two datasets: Eli5 provided by our mentor and the wiki40b_en_100_0 Wikipedia Snippets from hugging face.

The Eli5 dataset contains roughly 270,000 questions taken from the subreddit r/explainlikeimfive, along with answers with highest upvotes and their supporting context. Even though Reddit is home to many vibrant groups with excellent debates, it is also well recognized to have areas where misogyny, bigotry, and harassment are serious problems.

In order for our system to gather a large amount of knowledge for querying, we choose to give the model access to Wikipedia text. Wikipedia is well-known for it is home for millions of articles of numerous fields. However, full Wikipedia articles are too long for the majority of existing models to process. Thus, we use the wiki40b datasets with articles split into 100-word passages available on hugging face. Due to limited resources, we cannot process all 17 million entries of the wiki40b dataset. Therefore, we only use approximately 2,000,000 to 600,000 passages.

First, we embed all of the passages using the library sentence-transformer with bert-base-nli-mean-tokens pre-trained model from huggingface. For each passage the embedding has shape (1, 768). Then, we save our embedded data as vectors stored in PostgreSQL database in the form of id and data embedding.

For the Eli5 dataset, we use it for fine-tuning the BART model for generating answers. The model input format is as follows: [question_doc, answers], where variable question_doc is a text string with question plus all contexts relating to that question concatenated, and variable answers is all the answers to that question concatenated.

Fine-tuning BART for generating answers

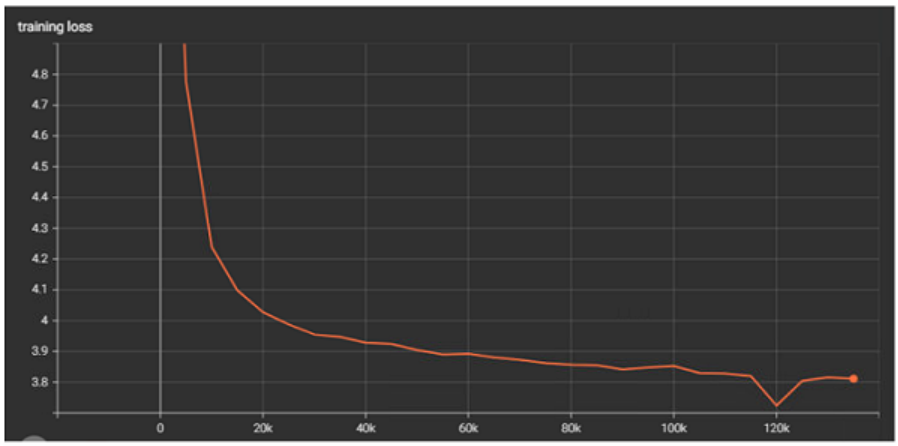
Now that our data is prepared, we move on to the next step, which is fine-tuning. The model we use for fine-tuning is the "yjernite/bart_eli5" pre-trained model from hugging face. The reason why we choose BART as our backbone for fine-tuning is because the BART model achieves state-of-the-art results on the challenging Eli5 dataset (Lewis et al, 2019).

The approach we use for fine-tuning the BART model is by applying LoRA. The LoRA method uses low-rank decomposition to divide the weight updates into two smaller matrices, known as update matrices. While minimising the overall number of changes, these new matrices may be taught to adjust to the new data. There are no further modifications made to the initial weight matrix, which stays static. The original and modified weights are combined to get the results. The LoRA configuration we use for our model is as follows: "r": 32, "lora_alpha": 32, "lora_dropout": 0.1, "target_modules": ["q_proj", "v_proj"], "task_type": "SEQ_2_SEQ_LM". We then apply this LoRA configuration to the BART backbone. Now, instead of training all 400 million parameters of the model, which is wasteful, we only train only 4 million parameters of our LoRA layers.

After setting up our LoRA layers, we then fine-tune our model for 3 epochs, with batch size of 2, learning rate of 2e-5, and using AdaFactor optimizer with default configurations. One epoch takes around 8 hours on a RTX 3060 6Gb vram, and we visualise the training losses after each epoch. After the training is finished, we then calculate the rougeL of our model using the validation split of the Eli5 dataset.

	Rouge1	Rouge2	RougeL
P	0.3097	0.0532	0.2777
R	0.2494	0.0469	0.2228
F	0.2455	0.0418	0.2191

Table 1: Rouge Evaluation of fine-tuned LoRA

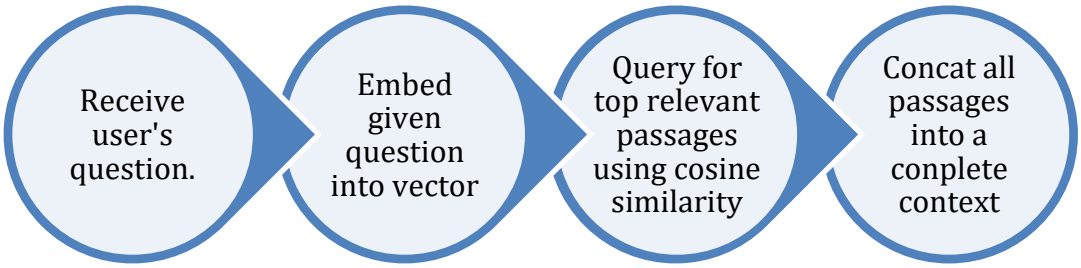


Training loss (record for every 5000 steps)

Creating context query function with sBERT

The next procedure in our process is to create our context query function. Thankfully, the sentence transformer library offers a selection of pre-trained models for embedding data. Hence, we can utilise one of the available models for our project. We decided on using the bert-base-nli-mean-token Sentence Transformer model, for it is lightweight and fast. The model is used for embedding the dataset and the input question.

For our database, we map each individual passage from the original wiki40b dataset, about 2,000,000 entries, into a 768-dimensional dense vector space. Then, we store our embedded dataset into PostgreSQL. The reason why we decided to use PostgreSQL is mainly because it supports storing data as vector format. Furthermore, Postgre also supports vector similarity search, which is the method we use to query top relevant passages for the given question.



The pipeline of context query function.

After we have our database up and running, as well as preparing the sBERT model for information retrieval, we then run some test examples to see whether our function works and evaluate our output. The question we use for testing derives from the test split of the Eli5 dataset, and we set k=5 for selecting top 5 most relevant passages based on the questions.

	Why can small terms to a power greater than 1 (i.e ^2, ^3 etc.) be neglected in many cases?
0	conjecture. Generalized Bunyakovsky conjecture Given k ≥ 1 polynomials with positive degrees and integer coefficients, each satisfying the three conditions, assume that for any prime p there is an n such that none of the values of the k polynomials at n are divisible by p. Given these assumptions, it is conjectured that there are infinitely many positive integers n such that all values of these k polynomials at x = n are prime. Note that the polynomials {x, x + 2, x + 4} do not satisfy the assumption, since one of their values must be divisible by 3
1	map (even just a rational map). For example, Spec k[x] and Spec k(x) and have the same function field (namely, k(x)) but there is no rational map from the former to the latter. However, it is true that any inclusion of function fields of algebraic varieties induces a dominant rational map (see morphism of algebraic varieties#Properties.)
2	it. Equivalently, the connectivity of a graph is the greatest integer k for which the graph is k-connected. While terminology varies, noun forms of connectedness-related properties often include the term connectivity. Thus, when discussing simply connected topological spaces, it is far more common to speak of simple connectivity than simple connectedness. On the other hand, in fields without a formally defined notion of connectivity, the word may be used as a synonym for connectedness. Another example of connectivity can be found in regular tilings. Here, the connectivity describes the number of neighbors accessible from a single tile:
3	R-modules if and only if in R, xy = 1 implies yx = 1. More generally, a Dedekind-finite ring is any ring that satisfies the latter condition. Beware that a ring may be Dedekind-finite even if its underlying set is Dedekind-infinite, e.g. the integers.
4	Irrelevant ideal In mathematics, the irrelevant ideal is the ideal of a graded ring generated by the homogeneous elements of degree greater than zero. More generally, a homogeneous ideal of a graded ring is called an irrelevant ideal if its radical contains the irrelevant ideal. The terminology arises from the connection with algebraic geometry. If $R = k[x_0, \ldots, x_n]$ (a multivariate polynomial ring in n+1 variables over an algebraically closed field k) graded with respect to degree, there is a bijective correspondence between projective algebraic sets in projective n-space over k and homogeneous, radical ideals of R not equal to the irrelevant ideal. More

	Why do people without restless legs syndrome shake their legs while seated?
0	consul of Russia in Haifa and the north of Israel. In 1991 he established Radio 1 (later become Haifa Radio) the first local radio station in Israel, broadcasting to the North and throughout the Haifa region. The Israeli Navy awarded Moshe Mano the title Notable of the Israeli Navy for his many years of contribution and assistance to both the sailors and the Navy in different domains including a memorial cruise to the INS Dakar submarine for the benefit of families of the crew. Mano also contributes to various volunteer associations and charitable organizations such as women's shelters, boarding schools, educational institutions for
1	Gry Forssell Early life and career Forssell grew up in Luleå and studied at the Child and youth education in high school. After graduating in 1992 she became television presenter for SVT's youth section in Växjö, along with Pernilla Månsson Colt and Per Dahlberg and they presented the shows PM and Pickup. She also acted in Ronny och Ragge as the character Bettan, a girl working the hot dog stand in Byhåla. Forssell is a hounorary member of Luleå Hockey and has a Luleå Hockey shirt hanging at her workplace studio at her radiostation Mix Megapol. After the work on SVT
2	be a teacher when the family moved to Brisbane, Queensland so she undertook commercial studies and obtained office work but with the birth of her children she developed an interest in dolls and toys. Personal life In 1953, Marjory married Jim Fainges. They had met at a friends' sister's 21st Birthday Party in late 1949, were engaged in 1951 and married in the Lutwyche Methodist Church, Brisbane. The family which was to include 5 children Lyn, Sue, Ian, Neil and Keith moved to Everton Park, Brisbane where Marjory and Jim worked on many projects together. Jim had his own
3	Boston University as an engineering graduate. While there he met his future wife, Constance (Cordice) Jackson, a teacher in the public school system. He and Constance later traveled to many places throughout the world and immersed themselves in the culture of every country they visited.
4	won equal honors at that university. Career The year following graduation was spent in teaching at a community school in Harmony, New Jersey, and then post-graduate work in the study of languages, particularly Latin, German, as well as music and drawing, and at the university commencement of 1875 she was given another degree, that of A. M. The young woman was, up to this time, planning to devote herself to teaching as a profession, but a visit to the Battle Creek Sanitarium in the summer of 1877 seems to have given an unconscious viewpoint to all her subsequent life. After teaching

Demo for context querying based on question.

It can be seen from the results that our model has successfully queried the passages from our database. However, the results vary widely. In some cases, the model can give out actual relevant passages, whereas in others the model returns irrelevant results. There are some factors that may contribute to this issue. For instance, the model we use is pre-trained on a different dataset. Additionally, our database only comprises a limited number of passages, hence it is not capable of covering all topics, which ultimately leads to the varying results.

Project Finalisation

This is the last step of our project, where we put all the pieces together into source code to push onto GitHub. In this stage, we compile everything we have done so far, from preparing the database, building the context query function as well as generating answers into one complete system. To help evaluate our project, we use Gradio, which is a library for creating UIs, to make a simple interface for our system and share it for evaluating. We collect 100 samples from various users and ask them to rate the answers from 1 to 5.

Question	Answer	Score
Bitrate vs Resolution?	Bitrate is the amount of data per second. Resolution is how much data is left. Bitrate = how many times the data is changed per second (or how many frames per second). Resolution = how much the data changes per second per second, or how many pixels per second the video is.	5
Why do airlines overbook their flights on purpose?	Because they know you're not going to book it on a regular basis. If you're booking it on the off chance you're going to be there, they'll overbook you. If they're booking you on the regular basis, they're not. They're booking the plane for you.	4
What happened to Gaddafi's money?	It was taken by the police and burned. The police took it to the bank and deposited it. The bank deposited it in the bank. It was then taken to the treasury and deposited there. There was no way for Gaddafi to have kept it. He was a dictator. He was a leader. He had no money. He didn't have any money.	1
What's so special about Counter Strike? It doesn't look very different from Modern Warfare. How does it draw such a dedicated community?	There's nothing special about Counter Strike. It's just a game. It's a game that you can play with your friends. It has a very limited skill set, but it's fun. It has a lot of fun. It is a game where you can be a badass. You can be an asshole. There is nothing special. The only thing special about it is that it is a sport. That's it. I'm not sure what else to say.	2

A sample of Question-Answer and the users' scores

After we collect the scores from 100 users, we estimate an average score of 2.5 out of 5[1], which is a decent score considering the finite resources in our opinion.

[1]: https://docs.google.com/spreadsheets/d/1g42EWqRr_1U9FnRiOcyYTTx0Eh9RuRKnH54mzFA29Jg/edit?usp=sharing

Conclusion

Artificial Intelligence in general, Natural Language Processing field in particular, are on the rise in recent years. Subsequently, the need for a system that can automate human tasks increases. Over the course of 2 weeks for the lab, we have successfully built an open domain long form question answering system that can generate answers for you by searching for the contexts required and give out long sentence answers based on the queried contexts.

This idea of a Question-Answering system has been around for a period of time now, and some companies have developed their very own system like OpenAI with ChatGPT, Microsoft with Bing Chat, etc. Our work is to implement the Transformers models and further enhance it for our purposes.

References

Yjernite (2022). Explain Anything Like I'm Five: A Model for Open Domain Long Form Question Answering : https://yjernite.github.io/lfqa.html?fbclid=IwAR1ohKu70O7EX0Q5jld3hTPozhQc-GpKRoYIyu0AbRXJcJ0eaEYgHlg2Gys#generation

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805v2.pdf

Hugging Face. https://huggingface.co/