Finding Similar Questions from ELI5 with Sentence-Bert

Ozan Yılmaz
Universität Heidelberg
Computerlinguistik
ozan.yilmaz@stud.uni-heidelberg.de

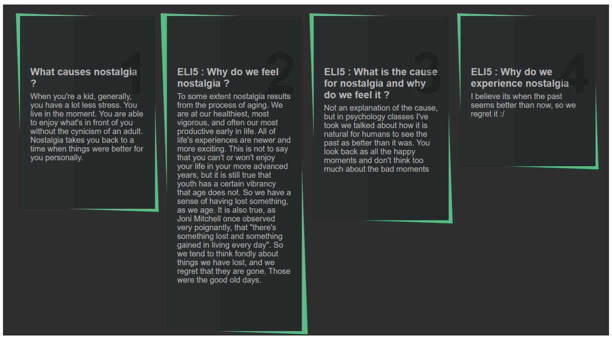
Abstract

This paper describes a semantic search application to find similar questions to a given query on the popular subreddit 'Explain Like I'm Five'. The system utilizes Sentence-Bert (SBERT) (4) by Reimers et al. 2019 to provide semantically meaningful results to the user. For every given query the full-fledged web-service returns the four semantically most similar questions that were asked by users of the subreddit in a timeframe between 2011 and 2018. Additionally, the system returns the best-rated answer to the question and a link to the original thread.
I use example questions to compare the results of SBERT with results from a TF-IDF Baseline to show a difference in quality. This system was designed as a proof-of-concept to show the ability of Sentence-Bert to be employed as a semantic search backend for customer support questions.

1 Introduction

It is essential for larger companies to reduce the workload of their customer support. Even though chatbots are gaining popularity, they are still far from perfect. A simpler but still very effective way to help customers is to provide them a guidance to find forum posts and answers to already asked similar questions before connecting them to a human help desk. Especially companies with large databases of already asked questions and help forums have enough data to provide a strategic search approach with the help of new NLP techniques.

I built a prototype of a 'similar asked questions' interface which shows question-answer result-pairs for a given question as input. As I don't have these kinds of data from companies, I built a proof-of-concept with the same idea using data from reddit's famous 'Explain like I'm 5' subreddit. Here, users can ask questions about nearly everything and receive layperson-friendly explanations. The system accepts a question and provides 4 similar questions, their best rated answers and the link to the original thread.



What causes nostalgia ?
When you're a kid, generally, you have a lot less stress. You live in the moment. You are able to enjoy what's in front of you without the cynicism of an adult. Nostalgia takes you back to a time when things were better for you personally.

ELI5 : Why do we feel nostalgia ?
To some extent nostalgia results from the process of aging. We are at our healthiest, most vigorous, and often our most productive early in life. All of life's experiences are newer and more exciting. This is not to say that you can't or won't enjoy your life in your more advanced years, but it is still true that youth has a certain vibrancy that age does not. So we have a sense of having lost something, as we age. It is also true, as Joni Mitchell once observed very poignantly, that "there's something lost and something gained in living every day". So we tend to think fondly about things we have lost, and we regret that they are gone. Those were the good old days.

ELI5 : What is the cause for nostalgia and why do we feel it ?
Not an explanation of the cause, but in psychology classes I've took we talked about how it is natural for humans to see the past as better than it was. You look back as all the happy moments and don't think too much about the bad moments

ELI5 : Why do we experience nostalgia
I believe its when the past seems better than now, so we regret it :/

Figure 1: The four most similar questions and their answers for the user-question 'Why do we feel nostalgia?'

2 Related Work

This application builds on two previous works:

Reimers et al.(2019)(4) present Sentence-BERT (SBERT), a modification of the original pretrained BERT network. SBERT can generate sentence embeddings, which makes it possible to calculate semantic textual similarity between sentences by utilizing measures like cosine-similarity. They use siamese network structures to derive semantically meaningful sentence embeddings. Their system outperforms other state-of-the-art sentence embedding methods in common STS tasks.
Facebook's project 'ELI5: Long Form Question Answering'(3) by Fan et al. 2019 builds a system that generates answers to open-domain questions. They extract questions and answers from Reddit's subreddit ELI5 and support documents from the

web to train and test their systems. Their complex seq2seq and language models generate good results for this hard task, although being still far from human standards.

This work combines ideas from these 2 papers by utilizing SBERT and the ELI5 dataset to create a system to find similar questions that were already asked by other users.

3 Background

3.1 BERT

BERT stands for Bidirectional Encoder Representations from Transformers and was introduced in 2018 by Devlin et al.(2) as a new pre-trained language representation model. The base system has a 12 layer architecture and is trained on masked language modeling and next-sentence prediction tasks with a 3.3B word English corpus. The fact that BERT is trained by jointly conditioning on both left and right context in all layers makes the resulting model even more powerful than previous language representation models. Fine-tuning the existing model with an additional output layer on top of it resulted in an improvement of the state-of-the-art for many relevant NLP tasks like question answering. The possibility to use BERT as a 'plug-and-play' system and finetune it with few iterations and a single layer on top for many tasks makes it especially useful for industry applications, where runtime and development time are very important resources and factors.

3.2 SBERT

For finding the two most similar sentences in a dateset, we would need to feed each unique pair through BERT. In a small dataset of only 10.000 sentences, this would require 49.995.000 passes through the system, which on a modern GPU would take 60+ hours. This renders BERT useless for STS-tasks from a computational point of view. Additionally, the vector space of BERT is not suitable for using distance measurements like cosine-similarity.

The authors of SBERT utilize twin networks and pool the outputs of 2 sentence inputs u and v to generate fixed-size vectors. They concatenate the output of the resulting vectors with the elementwise difference $\|u - v\|$ and multiply it with a trainable weight w (see Figure 2). They finetune this architecture on a combination of the SNLI(1) and the Multi-Genre NLI(5) corpus, which proved to be good data-sets for training sentence embeddings. Their results show major improvements in computational efficiency and outstanding results in STS-tasks, setting new state-of-the-art standards. SBERT reduces the computation time for the example which was mentioned in the beginning from 60+ hours to 5 seconds.

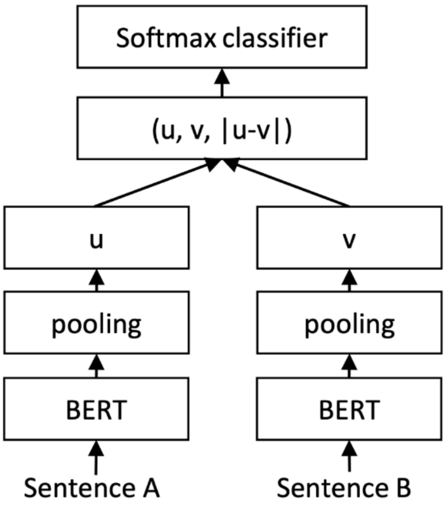


Figure 2: SBERT architecture with classification objective function

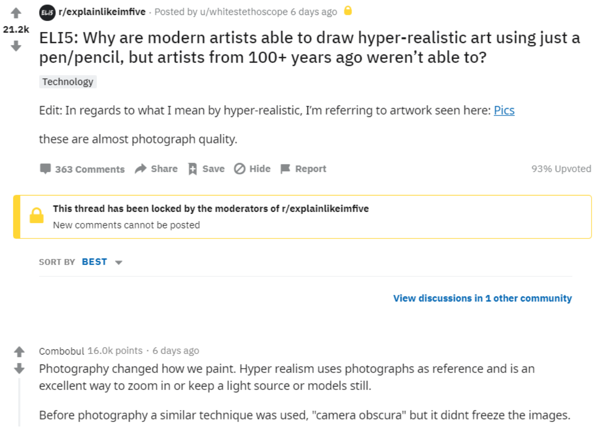
4 Data

I use Facebook's script 'download_reddit_qalist.py'[1] to download all threads from the subreddit 'Explain like I'm Five' which were posted between 2011 and 2018. A thread consists of a question which always starts with 'ELI5:', sometimes followed by a short text from the original poster and answers by other users (see Figure 3). We are interested in the question and the most voted answer, which should also be the most informative one.

After downloading these posts in a json format with the said script, I only extract the questions and links to the posts, which results in 254.608 separate queries. With this information, I create a seperate file. There is no further preprocessing involved, except for stop-word removal before TF-IDF vectorizing in the baseline model.

We have to keep in mind, that in a real customer-support scenario we would need to scrape or extract all these data with our own scripts and clean them up to make them usable, which results in much more work.

[1] https://github.com/facebookresearch/ELI5


21.2k
r/explainlikeimfive · Posted by u/whitestethoscope 6 days ago
Technology
363 Comments
Share
Save
Hide
Report
93% Upvoted
SORT BY BEST
View discussions in 1 other community
Combobul 16.0k points · 6 days ago

ELI5: Why are modern artists able to draw hyper-realistic art using just a pen/pencil, but artists from 100+ years ago weren't able to?

Edit: In regards to what I mean by hyper-realistic, I'm referring to artwork seen here: Pics

these are almost photograph quality.

This thread has been locked by the moderators of r/explainlikeimfive
New comments cannot be posted

Photography changed how we paint. Hyper realism uses photographs as reference and is an excellent way to zoom in or keep a light source or models still.

Before photography a similar technique was used, "camera obscura" but it didnt freeze the images.

Figure 3: An example thread in Reddit. We are interested in the question and the most voted answer.

5 Method

I encode all the extracted questions with the Sentence-Bert model 'bert-base-nli-mean-tokens' and save the encoded questions in a npy file. This allows me to quickly load the question vectors in production, encode the query and draw a comparison between the user question and the stack of pre-encoded reddit questions by using the cosine similarity measure. The resulting top four best fitting questions and their respective links are extracted with this method. As a next step, to get the most likely answer to the questions, I use the reddit API praw[2] to extract the top comment for each extracted link. This results in a set of four question-answer pairs, from which at least one hopefully answers the original question of the user.

It is important to keep in mind, that in a customer-support setting the domain would be more specialized and narrow. This could result in varying performances and we would have to evaluate if it makes sense to fine-tune SBERT with our data. For this step we would also need to analyze the quality of our data and discuss if we have enough to make an impact on the system by fine-tuning.

Evaluating an STS-system is a big task in itself, especially if there is no gold standard and the data is not annotated in one way or another. For this reason, I compare the differences between this sophisticated semantic search approach with a shallower TF-IDF approach by looking at different example sentences. Here, the idea is to pinpoint differences in results and explain why we can observe these.

TFIDF	SBERT
ELI5 : Why is there racism ?	ELI5 : Do animals express discrimination / racism based on the colour of fur / skin ?
ELI5 : Why is there racism ?	ELI5 : Do animals get embarrassed or feel shame ?
ELI5 : why is racism still a thing ?	ELI5 : Why is incest breeding bad for animals ?
ELI5 : Why is racism so common ?	ELI5 : Are there murderers , psychopaths or other behavioral deviations commonly associated with human beings among animals ?

Table 1: Resulting similar questions for the query 'Is there racism in animals?' sorted by cosine similarity

6 Results & Analysis

First of all, both systems provide meaningful results for common queries with common wordings. The difference between SBERT and TF-IDF becomes visible, when the user queries become more complex and semantically rich. TF-IDF struggles to decode the meaning behind such questions. These are the moments where SBERT shows it's potential. To highlight this behaviour, we will discuss some example queries in the following section.

The results in table 1 for the query 'Is there racism in animals' provides meaningful insights in how the SBERT system differs from the TF-IDF model. The top similar question for SBERT gives the user exactly what he asked for, although the wording of the question is different. Here we can observe the power of SBERT: The sentences are compared using a deeper level of semantics, which helps the system to 'understand' the query in a way which is not possible for a simple TF-IDF model. The baseline system gets very focused on the word 'racism' and provides two top answers containing 'is there racism'. It is obvious that the TF-IDF search mechanism drops the important phrase 'in

[2] https://github.com/praw-dev/praw

animals' in order to maximize the cosine-similarity. This makes it clear, that the baseline representation focuses on a word-level encoding structure.
Further, we can see that SBERT understands that racism is a 'bad feeling'. This is the reason why it lists other question results like 'Do animals get embarrassed or feel ashamed?'. This behaviour is out of scope for a TF-IDF model fitted on short questions.

TFIDF	SBERT
ELI5 : Could we see " someone " on Mars ?	ELI5 : why haven't we been able to land on Mars ?
ELI5 : Why not mars ?	ELI5 : Why haven't we put a man on Mars yet ?
ELI5 How high can a fly fly ?	ELI5 : Why are we making our expedition to mars a one way trip ?
ELI5 : Why does it sometimes cost more to fly from A → B than it does to fly from A → B → C ?	ELI5 : Why do we need to go to Mars ?

Table 2: Resulting similar questions for the query 'Why didn't we fly to Mars already?' sorted by cosine similarity

In the next example shown in table 2, we can see the results for the user question 'Why didn't we fly to Mars already?' The baseline model again concentrates more on the concepts of 'flying' and 'mars', while SBERT gets the intention of the question and provides the user with fitting results regarding the topic 'travel of mankind to mars'. It's interesting to observe the interchangeable usage of 'already' and 'yet' between the second SBERT-provided question and the user query. This again shows the semantic power of the BERT architecture.
The last example in table 3 asks the systems following question: 'Why do we need a social circle?'. The baseline system's first three questions completely focus on the word 'circle' and ignore the meaning of the phrase 'social circle'. The last answer on the list gives the user a satisfying result. SBERT on the other hand looks at the meaning of the whole sentence and realizes, that the user talks about human groups and social concepts. The system also takes the mental leap and understands, that the user means humans with 'we'.

These example sentences show the potential of SBERT to be used in a semantic search scenario. TF-IDF results would frustrate the customer in an automatic customer support setting, while SBERT would try to capture the whole semantics of the question and provide a meaningful answer, even for more obscure phrasings.
To give more insightful statements about the system, one would have to carry out systematic tests, including human evaluation steps.

TFIDF	SBERT
ELI5 : If you were to draw the universe as a circle on a piece of paper , what would be outside the circle ?	ELI5 : Can someone please explain to me what a social imaginary is ?
ELI5 : A circle is 360 . Is that arbitrary ? Could we divide a circle by 100 ?	ELI5 : Why do humans need social interaction ?
ELI5 : Can someone describe why a circle only has one dimension ?	ELI5 : What is a Caucus and why do we have them ?
ELI5 : Why do humans need social interaction ?	ELI5 : Why are humans wired for social connection ?

Table 3: Resulting similar questions for the query 'Why do we need a social circle?' sorted by cosine similarity

7 Conclusion

I built a semantic search application to find similar questions to a given query on the popular subreddit 'Explain Like I'm Five'. My system utilizes Sentence-Bert (4) by Reimers et al. 2019 to provide semantically meaningful results to the user. The intention of this project is to show the potential of SBERT in a customer-support setting.
First experiments display, that SBERT surpasses the TF-IDF baseline when it comes to understanding semantically more complex queries. It captures the meaning of the whole phrase and creates a satisfying output for the user. It could be possible, that the system would struggle in a more specialized domain like customer-support. If that would be the

case, one could fine-tune SBERT, given there is enough data to achieve the desired performance. There is need for further analysis of the system performance and results with labeled data and human evaluators and a concrete evaluation structure. All in all, I am confident that the system's performance is sufficient for the given task.

References

[1] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[2] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[3] Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190* (2019).

[4] Reimers, N., and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (11 2019), Association for Computational Linguistics.

[5] Williams, A., Nangia, N., and Bowman, S. R. The multi-genre nli corpus.