QA-Bert Explains it All:

Leveraging pre-trained transformers to rapidly build a Q-A system

Christopher J. Rico

Northwestern University

Introduction and Problem Statement

The world is inundated with information. Searching for and retrieving valid, specific information from a closed, unstructured corpus is a challenge for all those but the most practiced in NoSQL techniques. As the amount of available information grows and grows, methods for easier, natural-language information retrieval must be developed.

Enter Natural Language Processing. Since 2018, the artificial intelligence (AI) subfield of NLP has enjoyed rapid development thanks to advances in deep learning research and the advent of transfer learning techniques. Powerful pre-trained deep neural NLP models such as OpenAI's GPT, BERT and XLNet, developed by some of the biggest players in the AI world, have been made publically available. Leveraging these pre-trained neural models within a closed-domain question-and-answer system allows for rapid, high-level search and summarization of a corpus (Soares & Parreiras, 2018). Question-and-answer systems, which remain a challenge in the NLP world, offer a high-level, natural format in which to explore and retrieve specific information from a pre-structured database or a collection of natural language documents (a corpus). (Cao et al., 2010) consider QA systems an advanced form of information retrieval, one which has great utility because it delivers concise, question-specific answers.

This paper endeavors to explore and compare the efficacy of two different pre-trained NLP models (DistilBERT and RoBERTa) in the construction of a question-and-answer system to retrieve information from a corpus of movie reviews.

Literature Review

Question-and-answer systems are generally decomposed into three macro modules (Malik, Sharan, & Biswas, 2013, Bhoir & Potey, 2014): Question processing, Document

processing, and Answer processing. Question processing receives the input from the user and attempts to classify the type of question (think who, what, where, when, why, etc.). While manual classifications are typically rules-based, automatic classifications are more flexible and extensible to new question types (Ray, Singh, & Joshi, (2010). Document processing, the next module in the process, retrieves the most likely relevant documents or passages typically by performing a TF-IDF score on the corpus and matching it to the posed question (Malik et al., 2013). Answer processing is by far the most challenging part of the Q-A system, able to search through the documents retrieved by the previous module and extract a simple answer.

Dataset

The dataset consists of a single corpus of 61 text documents, each contributed by a student in the course. Within the corpus, each document is a film review written in English, truncated to roughly 500 words by the student who submitted it.

The Pre-trained models to be compared have been developed by two of the biggest players in the AI research space: Google and Facebook. BERT was developed by Google. Its name stands for Bidirectional Encoder Representations from Transformers. Unlike other recent language representation models, "BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers." In other words, BERT has the ability to consider sequences of words both forwards and backwards.There are many different versions of BERT. The two that will be compared in this paper are Facebook's RoBERTa (a heavily optimized version of BERT) (Liu, Ott, Goyal, Du, Joshi, Chen,. & Stoyanov, 2019) and HuggingFace's DistilBERT (a 'distilled' version of BERT) (Sanh, Debut, Chaumond & Wolf, 2019) Each of these are versions of BERT that have built on BERT's

method of language masking by tweaking the number of parameters and the size of the training set. Both models are versions which have been trained on the Stanford SQuAD dataset, a set of more than 100,000 human-answered question-and-answer pairs that are designed specifically for training NLP models to better comprehend text from a Q-A standpoint (Rajpurkar, Jia & Liang, 2018).

Both models (as well as many others) are publicly available at Huggingface.co/models.

Research Design and Methods

This experiment was executed in Google Colaboratory (a cloud-based Jupyter notebook) running a Python 3 interpreter. A variety of technologies were used to construct the Q-A system. To start, an Elasticsearch server acted as a repository for the corpus. From there, the open-source NLP library Haystack was used to construct the Q-A system. Haystack breaks a user-asked question down into its component parts, and allows for a quick TF-IDF or BM25 retrieval from the corpus to identify "candidate passages" that are likely to have the answer to the question. Then, the identified candidate passages are fed through the pretrained neural model to find answers in detail, returning the top n-answers, confidence scores, document where the answer was found, and the context passage where the answer was found.

To evaluate the system – and the choice of pretrained model – a series of questions about movies in the corpus were asked to the system. The correctness of the answer, as well as the confidence score of the answer, were compared between different models and varying difficulties of questions.

Results

Results are ordered by question, by model, in order of rank.

- Does Neo take the blue pill or the red pill?

RoBERTa

Answer	Document	Context	Probability	Score
blue	Matrix-1.txt	offered by Morpheus) or to return to his more normal “reality” (via the blu...	0.6117	3.6342
red	Martian_1.txt	Left for dead on the red planet following a scientifically anomalous but nar...	0.5727	2.3425
red	Dark_Knight_1.txt	mask for Ledger; his face is caked with moldy makeup that highlights the red...	0.4797	-0.6506
Ex Machina	Ex_Machina_3.txt	in Her pretty inevitable. But while Spike Jonze’s film was all heart, Ex Mac...	0.4199	-2.5853
Red	Star_Wars_2.txt	med with a literal staff, Rey has made it her life’s mission to part the Red...	0.3461	-5.0881

DistilBERT

Answer	Document	Context	Probability	Score
red	Matrix-1.txt	o’s choice is to embrace either the “really real” (as exemplified by the red...	0.7496	8.7728
Neo	Matrix-1.txt	ig-philosophical-questions-114007\nThe film centres on a computer hacker, “N...	0.7091	7.1277
blue one	Matrix-1.txt	ffered by Morpheus) or to return to his more normal “reality” (via the blue ...	0.7044	6.9466
Take the Joker	Dark_Knight_1.txt	cartoon fantasy). And the bad guys seem jazzed by their evildoing. Take the...	0.6191	3.8861
When the hundred guys come at Neo	Matrix_2.txt	Matrix “the Bible of the post-information age” and said, “When the hundred ...	0.4965	-0.1109

- When does Minority Report take place?

RoBERTa

Answer	Document	Context	Probability	Score
summer of 2002	Minority_Report_2.txt	ur computers in a similar way.\nWhen Minority Report came out in the summer ...	0.6781	5.9622
2002	Minority_Report_1.txt	to consider how rapidly technology has advanced since Minority Report's 200...	0.6258	4.1125
15th anniversary	Minority_Report_1.txt	We may have arrived at the 15th anniversary of Steven Spielberg's game-chang...	0.5312	1.0012
this summer	Minority_Report_3.txt	ould Spielberg have known the government would be using the same term this s...	0.4751	-0.7982
2015	Back_Future_3.txt	with the scuzzy plans of a guy named Griff, so McFly and Brown nip up to 201...	0.4475	-1.6847

DisitilBERT

Answer	Document	Context	Probability	Score
summer of 2002	Minority_Report_2.txt	ur computers in a similar way.\nWhen Minority Report came out in the summer ...	0.9156	19.0673
2002	Minority_Report_1.txt	to consider how rapidly technology has advanced since Minority Report's 200...	0.8520	14.0010
how can I do the cool computer swiping thing with my fingers	Minority_Report_1.txt	of criminal activity. \nA lot of people came away from Minority Report with...	0.7068	7.0398
2015	Back_Future_3.txt	erhaps even ours, on an even keel.\nIt all begins with danger in the year 20...	0.6621	5.3818
1985	Back_Future_3.txt	rld who loved the ride in Dr. Emmett Brown’s diabolical DeLorean back in 198...	0.6364	4.4793

- What did Matt Damon grow in The Martian?

RoBERTa

Answer	Document	Context	Probability	Score
Mr. Ripley	War_Games_2.txt	to have something to do with Matt Damon, because he was the Talented Mr. Ri...	0.5412	1.3206
corn	Interstellar_2.txt	nd his father-in-law (John Lithgow). Once a NASA pilot, Cooper now grows cor...	0.5159	0.5086
the ochre sands of Jordan’s Wadi Rum	Martian_1.txt	The Walk, The Martian keeps its visual palette positive, the ochre sands of ...	0.3486	-5.0003
older	Interstellar_2.txt	t. He and Murph remain on the ground, crunching the numbers and growing olde...	0.3264	-5.7949
Will Farmer	War_Games_1.txt	aren't tense/laughable? You'll be bored to tears. Matt Lanter plays Will Fa...	0.3171	-6.1381

DistilBERT

Answer	Document	Context	Probability	Score
corn	Interstellar_2.txt	nd his father-in-law (John Lithgow). Once a NASA pilot, Cooper now grows cor...	0.8997	17.5501
potatoes	Martian_1.txt	nson Crusoe situation, he discovers that it is indeed possible to grow potat...	0.8827	16.1428
computer hacker	War_Games_3.txt	e are still here. The hero, Will (Matt Lanter), is still a talented computer...	0.8412	13.3390
toys themselves start to grow up, even become parents	Toy_Story_1.txt	ble family, an unusually privileged one? And can toys themselves start to gr...	0.7335	8.1004
older	Interstellar_2.txt	t. He and Murph remain on the ground, crunching the numbers and growing olde...	0.6898	6.3946

- Who falls in love with Marty McFly?

RoBERTa

Answer	Document	Context	Probability	Score
Deckard	Blade_Runner_2.txt	an blond beast, Roy Batty (the wonderful Rutger Hauer). Along the way, Decka...	0.6023	3.3205
Jennifer	Back_Future_3.txt	an change history, or future history. With them is McFly’s girlfriend, Jenni...	0.5803	2.5908
Bonnie	Toy_Story_2.txt	t garbage can.\nIn the last few days before preschool begins in earnest, Bon...	0.5803	2.5906
Bonnie	Toy_Story_2.txt	anxiety even more to the fore.\nForky is a sorry specimen of a toy, but Bon...	0.5717	2.3098
Rachael	Blade_Runner_2.txt	Along the way, Deckard meets and falls in love with another replicant, Racha...	0.5374	1.2003

DistilBERT

Answer	Document	Context	Probability	Score
Deckard	Blade_Runner_2.txt	an blond beast, Roy Batty (the wonderful Rutger Hauer). Along the way, Decka...	0.9067	18.1933
Roy	Blade_Runner_1.txt	Deckard and between the pleasure model Priss (Daryl Hannah) and her lover Ro...	0.8550	14.1976
Michael J. Fox	Back_Future_3.txt	rty McFly’s son, Marty Jr. (conveniently played by McFly “himself,” Michael ...	0.8433	13.4640
Victoria	Cats_1.txt	tening inasmuch as the cast is remarkable. A shy and refined cat named Victo...	0.8271	12.5204
Harrison Ford	Blade_Runner_1.txt	When "Blade Runner" premiered in 1982, Harrison Ford disparagingly quipped, ...	0.8227	12.2756

- Who is Tony Stark’s nemesis?

RoBERTa

Answer	Document	Context	Probability	Score
Justin Hammer	Iron_Man_1.txt	proved only half true when Ivan is abducted by Stark's deadly enemy, Justin ...	0.8368	13.0779
Robert Downey Jr	Avengers_2.txt	t of course, this film is much more heavily populated. Tony Stark (Robert Do...	0.8064	11.4157
a father	Iron_Man_1.txt	anging over Tony Stark, as with most American heroes, is the shadow of a fat...	0.7688	9.6134
Doctor Strange	Avengers_1.txt	the alpha males have a tendency to bicker. Tony Stark is nettled by Doctor S...	0.7042	6.9408
Oliver Stone’s Wall Street	Big_Short_3.txt	on the apocalypse. Unlike the slick suits and killer sheen of Oliver Stone’s...	0.6889	6.3581

DistilBERT

Answer	Document	Context	Probability	Score
Justin Hammer	Iron_Man_1.txt	proved only half true when Ivan is abducted by Stark's deadly enemy, Justin ...	0.9452	22.7790
Robert Downey Jr	Avengers_2.txt	t of course, this film is much more heavily populated. Tony Stark (Robert Do...	0.9255	20.1599
Frodo Baggins	LOTR_1.txt	In THE LORD OF THE RINGS: THE FELLOWSHIP OF THE RING, our hero, Frodo Baggin...	0.9171	19.2269
Forky	Toy_Story_1.txt	and lovable material involves a beady-eyed Frankenstein’s monster named Fork...	0.9075	18.2697
Doctor Strange	Avengers_1.txt	the alpha males have a tendency to bicker. Tony Stark is nettled by Doctor S...	0.9037	17.9131

The Q-A system using RoBERTa was able to return 9 of 30 questions correctly (precision score of 0.6) with an average confidence score of 2.739, while the system using DistilBERT was able to return 8 of 30 questions correctly (precision score of 0.5333) with an average confidence score of 12.248.

Analysis

Qualitative Analysis

Overall, Q-A systems using the two pretrained models performed very similarly in terms of correct answers returned. Each system managed to return at least one correct answer in n-answers per question, besides the very tricky question of "Who falls in love with Marty McFly?". However, the RoBERTa-backed system consistently returned answers with a much lower confidence score than the DistilBERT-backed system. This may be an artifact of the greater number of parameters in RoBERTa model: as the potential documents are subject to greater scrutiny by a larger number of parameters, the model is less able to return an answer with certainty. The often ironically stated Segal's law states that "someone with one watch is sure of the time, while someone with two watches is not so sure". While that's somewhat of a paradox, it's interesting to see that a larger model returns answers with less certainty. Perhaps even neural nets are subject to the Dunning-Kruger effect.

Qualitative Analysis

What is most interesting about the Q-A system is the way that it occasionally returns correct answers that refer to the incorrect film. For example, when asked if "Neo took the red pill or the blue pill?", the RoBERTa-backed system returned the answer "red" as a second guess. The reference document, though, was a review of the film The Martian -- not even remotely close to the actual film that Neo stars in, The Matrix.

Because each of these pre-trained models are already optimized for Q-A systems, they are both fairly adept at understanding the basic structure of a question: in the aforementioned example, the system was able to differentiate that either "red" or "blue" was an acceptable answer, and it

scavenged the corpus for confirmatory passages and chose several, even though some of those passages were straight-up from the wrong movie.

On the other hand, some of the answers that the systems return are outright poetic, especially when one considers the passage context that influenced the decision. Take, for example, the question “Who is Tony Stark’s nemesis?”. An acceptable answer would have been “Thanos”, “Justin Hammer”, “Loki”, or any of the other antagonist characters from the Iron Man or Avengers films. Indeed, the systems return these answers, but they also return answers such as “Oliver Stone’s Wall Street”, and “a father”. The words “killer” and “shadow”, which are both located close to the answers, appear to influence the model’s decision to output these answers. Another example of an unintentionally poetic answer is one of the responses to “What does Matt Damon grow in The Martian?”. While only the DistilBERT-backed system correctly answers “potatoes”, both systems return the response of “older”, which is a beautifully abstract way of looking at the passage of time.

Technology Considerations

The scope of this project ended up being smaller than I had originally intended. At first, I planned on building a COVID-19 Q-A system from the ground up using only the term extraction and vectorization techniques learned in this class. After what was probably far too much tinkering trying to engineer the system, I realized that doing this without any help from outside libraries was probably not going to be a possibility without some serious time invested. I settled on simply building a Q-A system and assessing a few pretrained models on a corpus that I was familiar with. Building a Q-A system for the COVID-19 research corpus will be a project for over the summer.

Conclusions

In conclusion, I was able to successfully build a Q-A system that leveraged a pre-trained NLP model. While the answer results for each pre-trained model were roughly similar, both systems struggled somewhat with more abstract questions, and occasionally even selected the right answer from the wrong document!

This was a very interesting look into the inner workings of Q-A systems. While they simplify and make more abstract the indexing and retrieval of information, their answers can be somewhat ambiguous, needing further review from a human before trusting the results. With further refinement, a system like this could be applied to a body of research like the COVID-19 corpus to help researchers learn from an otherwise impossibly large corpus.

References

Soares, M. A. C., & Parreiras, F. S. (2018). A literature review on question answering techniques, paradigms and systems. Journal of King Saud University-Computer and Information Sciences.

Cao, Y. G., Cimino, J. J., Ely, J., & Yu, H. (2010). Automatically extracting information needs from complex clinical questions. Journal of biomedical informatics, 43(6), 962-971.

Malik, N., Sharan, A., & Biswas, P. (2013, December). Domain knowledge enriched framework for restricted domain question answering system. In 2013 IEEE International Conference on Computational Intelligence and Computing Research (pp. 1-7). IEEE.

Bhoir, V., & Potey, M. A. (2014, February). Question answering system: A heuristic approach. In The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014) (pp. 165-170). IEEE.

Ray, S. K., Singh, S., & Joshi, B. P. (2010). A semantic approach for question classification using WordNet and Wikipedia. Pattern Recognition Letters, 31(13), 1935-1943.

Yao, X. (2014). Feature-driven question answering with natural language alignment (Doctoral dissertation, Johns Hopkins University).

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822.