Question Answering using Transformers

Shubham Nishad
300102830
snish035@uottawa.ca

CSI5180 Topics in AI: Virtual Assistants

Project summary

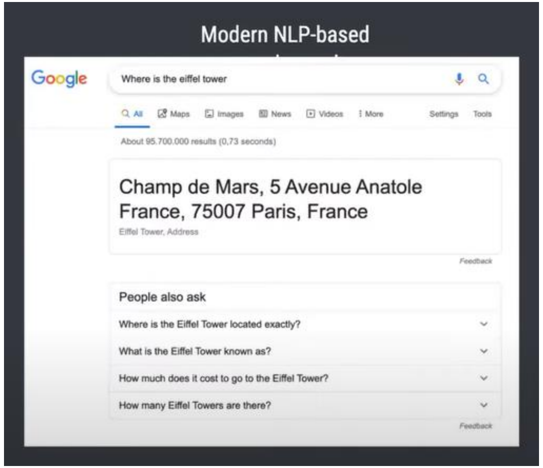
Virtual Assistants → Question Answering → Extractive QA → Supervised Deep Learning

What? Comparison study on latest transformer models and their use in Question Answering.

Why QA? Transformer based models have matched human evaluation and allowed for practical usage.

Reason for this project? Big momentum in research, Hot topic in NLP

Final Deliverable: Deep Learning system capable of extractive QA.



Modern NLP-based

Where is the eiffel tower

About 95.700.000 results (0,73 seconds)

Champ de Mars, 5 Avenue Anatole France, 75007 Paris, France

Eiffel Tower, Address

Feedback

People also ask

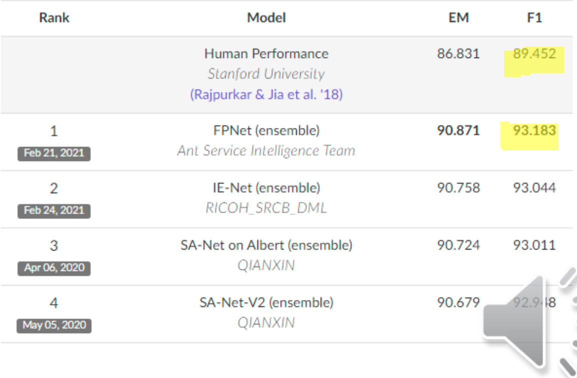
Where is the Eiffel Tower located exactly?

What is the Eiffel Tower known as?

How much does it cost to go to the Eiffel Tower?

How many Eiffel Towers are there?

Feedback



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	**90.871**	**93.183**
2 Feb 24, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	90.758	93.044
3 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
4 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948

Resources used

Transformers Library : The goto library for NLP. Has pre trained state of the art models and various standard datasets.

Google Colaboratory : Free GPU support, configured for DL, easy code sharing

Stanford Question Answering Dataset (SQuAD) : standard dataset for extractive QA.

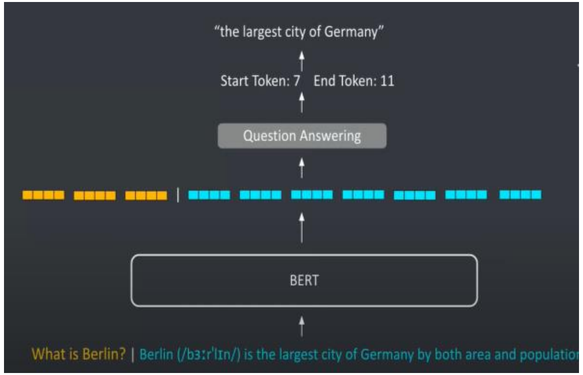
COVID-QA : Dataset developed to answer questions related to COVID-19.

Github link : Code repository for this project

Methodology

Popular Approach

1. Train Model to identify the start and end of answer span.
2. Fine-tune a language model on labelled question-answer pairs.
3. SQuAD has become a default dataset.

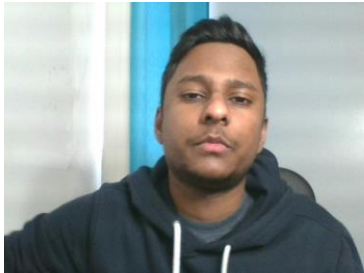


What is Berlin? | Berlin (/bɜːrˈlɪn/) is the largest city of Germany by both area and populatio

The Process

1. Considered the SQuAD dataset leaderboard and Transformers library. Selected 5 common language models.

- BERT
- DistilBERT
- ELECTRA
- MobileBert
- RoBERTa



Methodology

2. Train the models on SQuAD dataset, experiment to find best parameters.

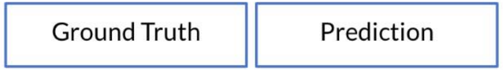
3. Train the models on COVID-QA dataset, experiment to find best parameters.

4. Compare models using two metrics: **Exact Match (EM)** and **F1 Score**.

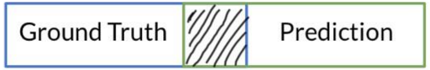
5. Perform statistical analysis on the datasets and correlate with performance.

6. Try to find concrete reasons for irregularities and special cases.

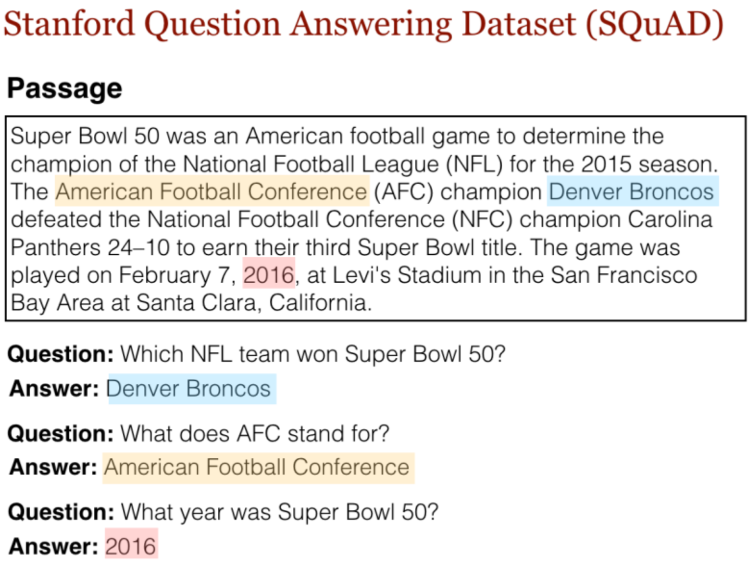
Exact Match: predicted exactly matches the ground truth



F1: measure of how close the prediction is to the ground truth



Datasets



Stanford Question Answering Dataset (SQuAD)

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

Question: Which NFL team won Super Bowl 50?
Answer: Denver Broncos

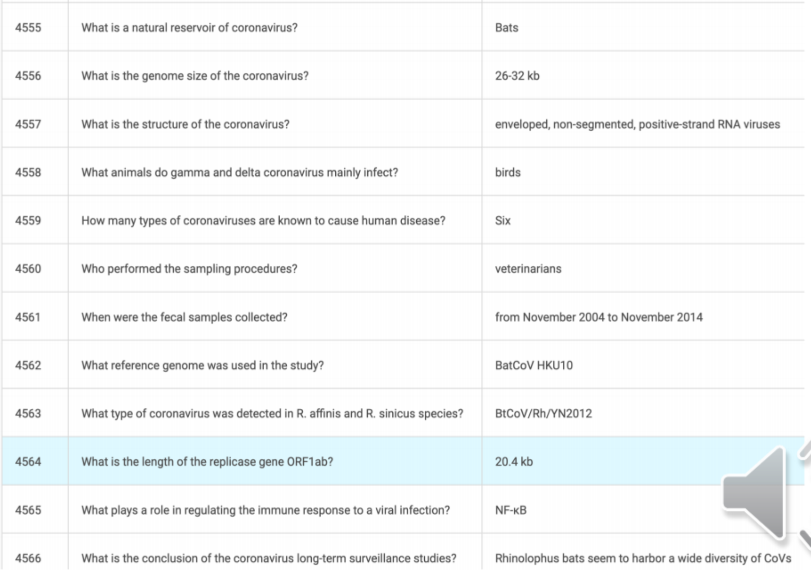
Question: What does AFC stand for?
Answer: American Football Conference

Question: What year was Super Bowl 50?
Answer: 2016

107,785 question-answer pairs on 536 articles.

COVID-QA

2,019 Question Answering pairs from 147 long articles.



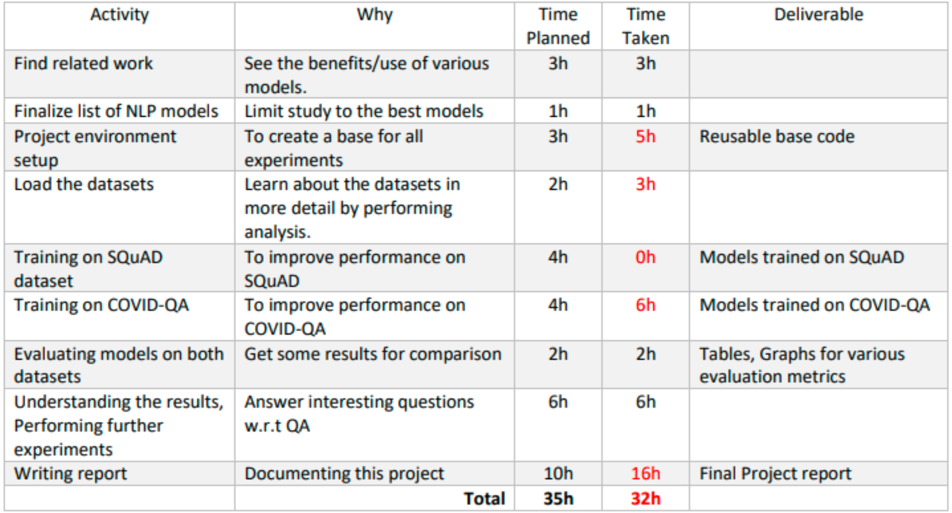
4555	What is a natural reservoir of coronavirus?	Bats
4556	What is the genome size of the coronavirus?	26-32 kb
4557	What is the structure of the coronavirus?	enveloped, non-segmented, positive-strand RNA viruses
4558	What animals do gamma and delta coronavirus mainly infect?	birds
4559	How many types of coronaviruses are known to cause human disease?	Six
4560	Who performed the sampling procedures?	veterinarians
4561	When were the fecal samples collected?	from November 2004 to November 2014
4562	What reference genome was used in the study?	BatCoV HKU10
4563	What type of coronavirus was detected in R. affinis and R. sinicus species?	BtCoV/Rh/YN2012
4564	What is the length of the replicase gene ORF1ab?	20.4 kb
4565	What plays a role in regulating the immune response to a viral infection?	NF-κB
4566	What is the conclusion of the coronavirus long-term surveillance studies?	Rhinolophus bats seem to harbor a wide diversity of CoVs

Activity table

Writing modular base code and loading datasets took 60% more time than estimate.

Able to find pre-trained models because of the popularity of SQuAD dataset.

Training was time consuming, ran tasks in parallel.

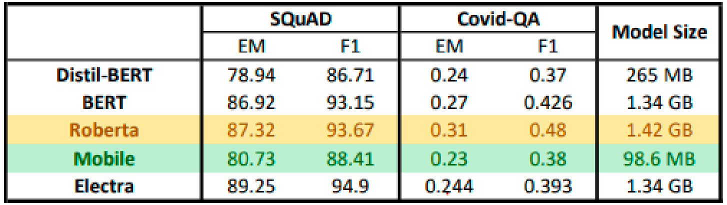


Activity	Why	Time Planned	Time Taken	Deliverable
Find related work	See the benefits/use of various models.	3h	3h	
Finalize list of NLP models	Limit study to the best models	1h	1h	
Project environment setup	To create a base for all experiments	3h	5h	Reusable base code
Load the datasets	Learn about the datasets in more detail by performing analysis.	2h	3h	
Training on SQuAD dataset	To improve performance on SQuAD	4h	0h	Models trained on SQuAD
Training on COVID-QA	To improve performance on COVID-QA	4h	6h	Models trained on COVID-QA
Evaluating models on both datasets	Get some results for comparison	2h	2h	Tables, Graphs for various evaluation metrics
Understanding the results, Performing further experiments	Answer interesting questions w.r.t QA	6h	6h	
Writing report	Documenting this project	10h	16h	Final Project report
	Total	**35h**	**32h**	

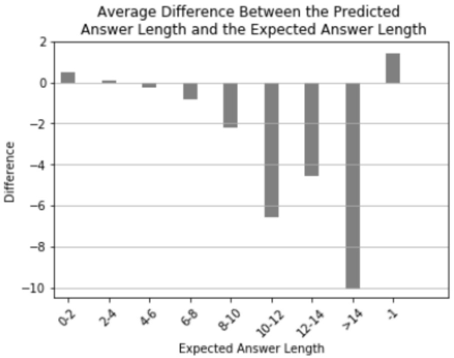
Results

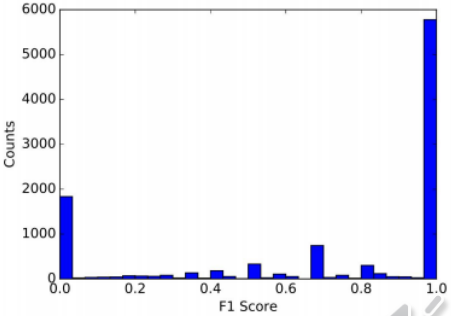
Comparison of Models

F1 and EM scores are low compared to metrics reported on SQuAD because of the different text domain as well as an about 40x larger text size.



	SQuAD		Covid-QA		Model Size
	EM	F1	EM	F1	
Distil-BERT	78.94	86.71	0.24	0.37	265 MB
BERT	86.92	93.15	0.27	0.426	1.34 GB
Roberta	87.32	93.67	0.31	0.48	1.42 GB
Mobile	80.73	88.41	0.23	0.38	98.6 MB
Electra	89.25	94.9	0.244	0.393	1.34 GB





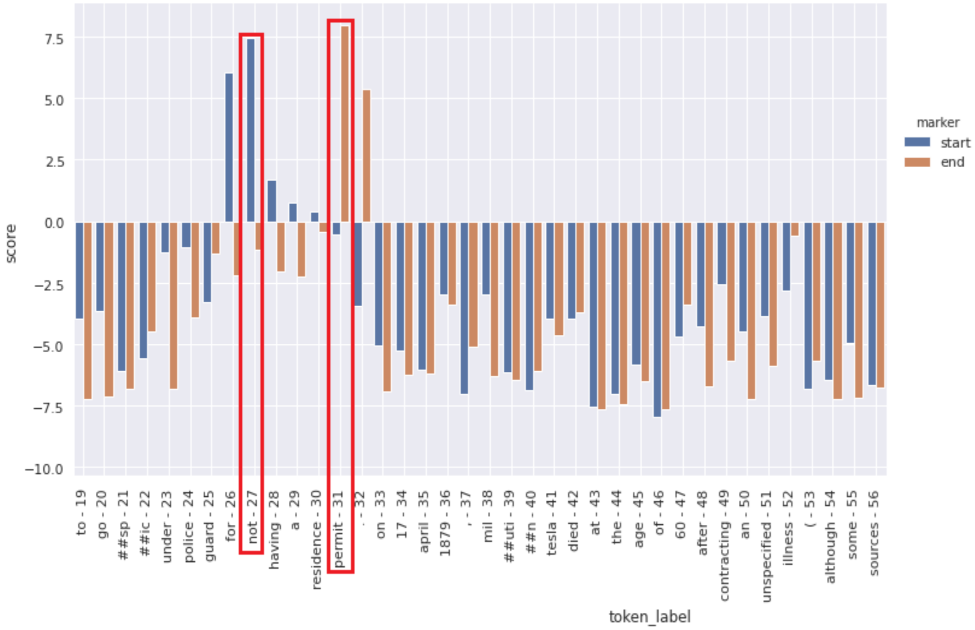
Results

Attention Mechanism

BERT Base has 12 layers and 12 heads, resulting in a total of 12 x 12 = 144 distinct attention mechanisms.

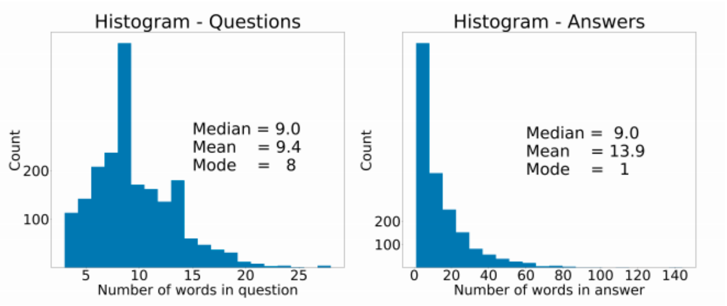
Attention heads do not share parameters, each head learns a unique attention pattern.

From our study it can be said that attention mechanism being used is more important than the underlying architecture. Supported by (Vaswani et al., 2017).

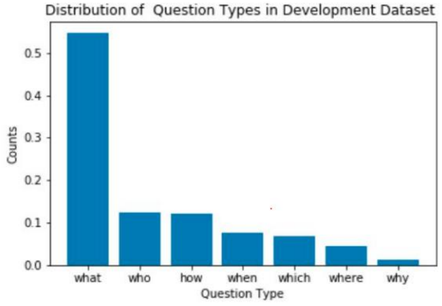


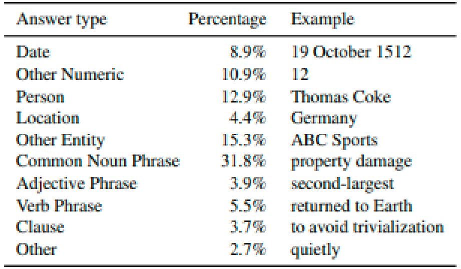
Results

Statistical analysis of Datasets

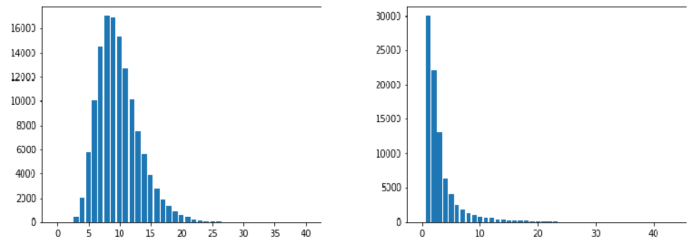


Histogram for COVID-QA





Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly



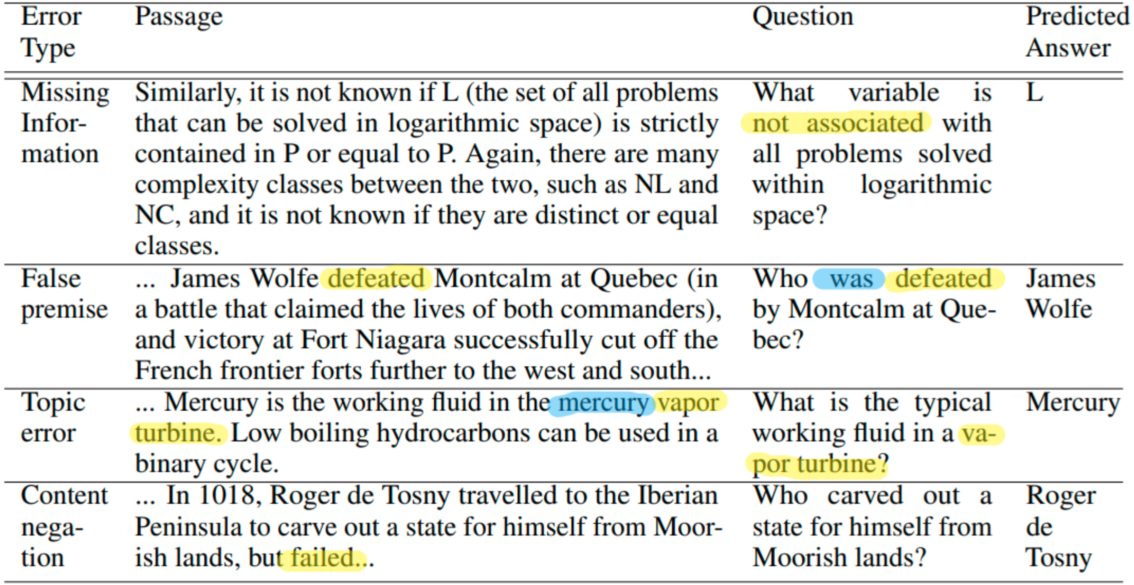
Question Length

Answer Length

Histogram for SQuAD

Results

Error Analysis



Error Type	Passage	Question	Predicted Answer
Missing Information	Similarly, it is not known if L (the set of all problems that can be solved in logarithmic space) is strictly contained in P or equal to P. Again, there are many complexity classes between the two, such as NL and NC, and it is not known if they are distinct or equal classes.	What variable is not associated with all problems solved within logarithmic space?	L
False premise	... James Wolfe defeated Montcalm at Quebec (in a battle that claimed the lives of both commanders), and victory at Fort Niagara successfully cut off the French frontier forts further to the west and south...	Who was defeated by Montcalm at Quebec?	James Wolfe
Topic error	... Mercury is the working fluid in the mercury vapor turbine. Low boiling hydrocarbons can be used in a binary cycle.	What is the typical working fluid in a vapor turbine?	Mercury
Content negation	... In 1018, Roger de Tosny travelled to the Iberian Peninsula to carve out a state for himself from Moorish lands, but failed...	Who carved out a state for himself from Moorish lands?	Roger de Tosny

source

Challenges

Challenges faced during development of this project

1. Limited Computation Power and Long training time.
2. Getting the COVID-QA dataset to work with the code built for SQuAD dataset.
3. Performing manual and statistical comparison of the results.

Challenges of Question Answering

1. Interaction of text and question
2. Large set of potential predictions. O(n^2)
3. Contextual understanding of potentially long text

What have you learned?

1. Transformers library for developing NLP related projects.
2. Process of designing end to end Question Answering Systems using Deep Learning.
3. The basics of multiple transformer based Language Models.
4. Various dataset available for QA. The different domains and their effect on performance.
5. Importance of Attention Mechanism in NLP.

Conclusion

- Successfully used and compared multiple production ready model on two different datasets.
- Learned the process of designing a QA system using NLP.
- Found some interesting observations that would help in future work.
- Saw the importance of dataset quality and model fine tuning on the performance.

References

1. Question Answering Beyond SQuAD: Larger Datasets and New Domains, with Branden Chan, deepset.ai https://www.youtube.com/watch?v=E80qHThomok

2. Applying BERT to Question Answering (SQuAD v1.1) https://www.youtube.com/watch?v=l8ZYCvgGu0o

3. BertViz https://github.com/jessevig/bertviz

4. Question Answering using Google's Natural Question Dataset and BERT - Full Summary https://www.youtube.com/watch?v=d_TsJ4IjlQQ