

# Tri-Lingual Question- Answering System

---

Nithin Chandra Gupta Samudrala  
200110076

# Motivation

---

- A trilingual question answering system provides students with the ability to ask questions and gain a deeper understanding of content across multiple languages.

# Problem Statement

---

## **Building Question-Answering System in Telugu, English and Hindi.**

For each observation in the training set we have a context, question, and answer.

The goal is to find the answer for any new question and context provided.

I used “Extractive Question Answering” Method. This involves posing questions about a document and identifying the answers as spans of text in the document itself. This makes it a closed dataset meaning that the answer to a question is always a part of the context and also a continuous span of context.

# Problem Statement - Contd.

---

## **Building Question-Answering System in Telugu, English and Hindi.**

The context can be in any of the three languages: Telugu, English and Hindi.

The question can also be in any of the three languages mentioned above.

I used google-translate API to translate any text to English before passing it to our model.

# Literature Survey

---

**Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang:**

SQuAD: 100,000+ Questions for Machine Comprehension of Text

[Link to paper](#)

# Dataset

---

Stanford Question Answering Dataset (**SQuAD**) is a new reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage. With 100,000+ question-answer pairs on 500+ articles, SQuAD is significantly larger than previous reading comprehension datasets.

# Dataset Contd.

Loading the Dataset:

```
from datasets import load_dataset  
  
raw_datasets = load_dataset("squad")
```

print(raw\_datasets):

```
DatasetDict({  
  train: Dataset({  
    features: ['id', 'title', 'context', 'question', 'answers'],  
    num_rows: 87599  
  })  
  validation: Dataset({  
    features: ['id', 'title', 'context', 'question', 'answers'],  
    num_rows: 10570  
  })  
})
```

# Dataset Contd.

The training dataset has only one possible answer while the validation dataset has several possible answers, which may be same or different

```
raw_datasets["train"].filter(lambda x: len(x["answers"]["text"]) != 1)
```

```
Dataset({
  features: ['id', 'title', 'context', 'question', 'answers'],
  num_rows: 0
})
```

```
print(raw_datasets["validation"][0]["answers"])
print(raw_datasets["validation"][2]["answers"])
```

```
{'text': ['Denver Broncos', 'Denver Broncos', 'Denver Broncos'], 'answer_start': [177, 177, 177]}
```

```
{'text': ['Santa Clara, California', 'Levi's Stadium', 'Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.'], 'answer_start': [403, 355, 355]}
```



# Model

- I have used a BERT Model (Bidirectional Encoder Representations from Transformers) and fine-tuned it for our task.
- The Architecture of this model is shown in the picture on the right-side.

```
BertModel(  
  (embeddings): BertEmbeddings(  
    (word_embeddings): Embedding(28996, 768, padding_idx=0)  
    (position_embeddings): Embedding(512, 768)  
    (token_type_embeddings): Embedding(2, 768)  
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)  
    (dropout): Dropout(p=0.1, inplace=False)  
  )  
  (encoder): BertEncoder(  
    (layer): ModuleList(  
      (0-11): 12 x BertLayer(  
        (attention): BertAttention(  
          (self): BertSelfAttention(  
            (query): Linear(in_features=768, out_features=768, bias=True)  
            (key): Linear(in_features=768, out_features=768, bias=True)  
            (value): Linear(in_features=768, out_features=768, bias=True)  
            (dropout): Dropout(p=0.1, inplace=False)  
          )  
          (output): BertSelfOutput(  
            (dense): Linear(in_features=768, out_features=768, bias=True)  
            (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)  
            (dropout): Dropout(p=0.1, inplace=False)  
          )  
        )  
      )  
      (intermediate): BertIntermediate(  
        (dense): Linear(in_features=768, out_features=3072, bias=True)  
        (intermediate_act_fn): GELUActivation()  
      )  
      (output): BertOutput(  
        (dense): Linear(in_features=3072, out_features=768, bias=True)  
        (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)  
        (dropout): Dropout(p=0.1, inplace=False)  
      )  
    )  
  )  
  (pooler): BertPooler(  
    (dense): Linear(in_features=768, out_features=768, bias=True)  
    (activation): Tanh()  
  )  
)
```

# Metrics

---

- **Exact match** : This metric calculates exact similarity between predict answer and the ground truth. For question answering, it serves as a good metric as we need answer that has to be accurate to question asked
- **f1** : it is the harmonic mean of the precision and recall of the predicted answer and ground truth answer.

# Results

```
predictions = model.predict(tf_eval_dataset)
compute_metrics(
    predictions["start_logits"],
    predictions["end_logits"],
    validation_dataset,
    raw_datasets["validation"],
)
```

677/677 [=====] - 338s 495ms/step

100%  10570/10570 [00:20<00:00, 380.13it/s]

{'exact\_match': 80.45411542100284, 'f1': 88.26667957260283}

# Results and Analysis

- BERT worked pretty well in extracting answers for factoid questions like “Who is the current captain of CSK?” but fared poorly when given open-ended questions such as “Is Dhoni the best captain?”.
- This is because we trained the model on dataset which has fact-based questions with explicit answers. To answer open-ended questions, it needs better interpretation and understanding.
- Answering open-ended questions might also come under the category of generative tasks which is tough for BERT as it's language and concepts of relations between words is limited to the training data and also due to its MLM objective.
- Although MLM objective is useful for learning representations of words and their relationships with other objects, it is not as effective for learning how to generate grammatically correct text.
- When model is asked out of context question, model is hallucinating which is undesired

# Streamlit Question Answering App

Enter Context:

మహేంద్ర సింగ్ ధోని భారత మాజీ అంతర్జాతీయ క్రికెటర్, అతను 2007 నుండి 2017 వరకు పరమేత ఓపెన్ బ్యాట్స్ మన్ మరియు 2008 నుండి 2014 వరకు టెస్ట్ క్రికెట్ లో భారత జాతీయ క్రికెట్ జట్టుకు క్యాప్టెన్ గా ఉన్నాడు మరియు ప్రస్తుత CSK కెప్టెన్. అతను మూడు ICC ట్రోఫీలలో భారతీయ విజయశీరాసకు చేర్చాడు, ఏ భారత కెప్టెన్ లోనూ అత్యధిక విజయాన్ని సాధించాడు. భారత టెస్ట్ క్రికెట్ లో అతను టెస్ట్ మరియు ట్వంటీ ట్వంటీ క్రికెట్ జట్టుకు ఆదాడు. ఇండియన్ ప్రీమియర్ లీగ్ లో వెస్ట్ సూపర్ కింగ్స్ (CSK) కెప్టెన్ గా ఉన్నాడు. అతను IPL లీగ్ యొక్క 2010, 2011, 2018 మరియు 2021 ఎడిషన్లలో ఛాంపియన్ షిప్ లకు నాయకత్వం వహించాడు. అతని కెప్టెన్సీలో వెస్ట్ సూపర్ కింగ్స్ (CSK) 2010 మరియు 2014లో రెండు సార్లు ఛాంపియన్స్ లీగ్ T20ని గెలుచుకుంది. ధోని తన ODI అరంగేట్రం 23 డిసెంబర్ 2004న, బంగ్లాదేశ్ పై టెస్టుగా లో, [2] మరియు ఒక సంవత్సరం తర్వాత ఫ్రీట్ ఆడాడు. తన [3] అతను ఒక సంవత్సరం తర్వాత టెస్ట్ కెప్టెన్ గా తన మొదటి T20 గూడా ఆడాడు. [4] 2007లో, అతను రెండవ ప్రపంచ కప్ ODI కెప్టెన్ గా నిర్ణయించాడు మరియు ఈ సంవత్సరంలో అతను భారతదేశానికి T20 కెప్టెన్ గా కూడా ఎంపికయ్యాడు. [5] 2008లో, అతను టెస్ట్ కెప్టెన్ గా ఎంపికయ్యాడు. [6] టెస్ట్ ఫార్మాట్ లో అతని కెప్టెన్సీ రికార్డు మిశ్రమంగా ఉంది, 2008లో న్యూజీలాండ్ పై సిరీస్ విజయం సాధించి, బోర్న్-గవాస్కర్ ట్రోఫీ (2010 మరియు 2013లో విజయ్ సిరీస్) అక్వీరియ్యూ ఫ్రీలండ్, అఫ్ఘనిస్తాన్, ఇంగ్లాండ్ మరియు దక్షిణాఫ్రికా చేతిలో ఓడిపోయింది. దూరంగా ఉన్న పర్యటనల్లో పెద్ద మార్గదర్శకత్వం. [7] అతను 30 డిసెంబర్ 2014న టెస్ట్ ఫార్మాట్ నుండి రిటైర్మెంట్ ప్రకటించాడు. [8] మరియు 2017లో T20లు మరియు ODIల కెప్టెన్సీ నుండి వైదొలగాడు

Select Context Language:

Telugu

Enter Question:

in which city did dhoni make his odi debut

Select Question Language:

English

Predict Answer

english

Answer: Chittagong

# Results

Context-Telugu  
Question-English

# Streamlit Question Answering App

Enter Context:

మహేంద్ర సెంగ్ ధోని భారత మాజీ అంతర్జాతీయ క్రికెటర్, అతను 2007 నుండి 2017 వరకు పరిమిత ఓవర్ల ఫార్మాట్‌లలో మరియు 2008 నుండి 2014 వరకు టెస్ట్ క్రికెట్‌లో భారత జాతీయ క్రికెట్ జట్టుకు కెప్టెన్‌గా ఉన్నారు మరియు ప్రస్తుత CSK కెప్టెన్. అతను మూడు ICC ట్రోఫీలలో భారతను విజయశీరాసకు చేర్చాడు, ఏ భారత కెప్టెన్‌లోనూ అత్యధిక విజయాన్ని సాధించాడు. భారత టెస్ట్ క్రికెట్‌లో అతను టెస్ట్ మరియు జూనియర్ క్రికెట్ జట్లకు ఆటాడు. ఇండియన్ ప్రీమియర్ లీగ్‌లో చెన్నై సూపర్ కింగ్స్ (CSK) కెప్టెన్‌గా ఉన్నారు. అతను IPL లీగ్ యొక్క 2010, 2011, 2018 మరియు 2021 ఎడిషన్లలో ఛాంపియన్‌షిప్‌లకు వాయిదాతర్ఫు వహించాడు. అతని కెప్టెన్సీలో చెన్నై సూపర్ కింగ్స్ (CSK) 2010 మరియు 2014లో రెండు వార్డు ఛాంపియన్స్ లీగ్ T20ని గెలుచుకుంది. ధోని తన ODI అరంగేట్రం 23 డిసెంబర్ 2004న, బంగ్లాదేశ్‌పై ఫిట్టాంగ్‌లో,[2] మరియు 48 సంవత్సరం తర్వాత శ్రీలో ఆడాడు. తండ్రి.[3] అతను 48 సంవత్సరం తర్వాత దక్షిణాఫ్రికాపై తన మొదటి T20I కూడా ఆడాడు.[4] 2007లో, అతను లాచెంట్ ట్రవెల్ నుండి ODI కెప్టెన్సీని స్వీకరించాడు మరియు ఈ సంవత్సరంలో అతను భారతదేశానికి T20I కెప్టెన్‌గా కూడా ఎంపికయ్యాడు.[5] 2008లో, అతను టెస్ట్ కెప్టెన్‌గా ఎంపికయ్యాడు.[6] టెస్ట్ ఫార్మాట్‌లో అతని కెప్టెన్సీ రికార్డు మిశ్రమంగా ఉంది, 2008లో స్కాట్లాండ్‌పై సిరీస్ విజయం సాధించి, బోర్డర్-గవాస్కర్ ట్రోఫీ (2010 మరియు 2013లో స్వదేశీ సిరీస్) ఆస్ట్రేలియాపై శ్రీలంక, ఆఫ్ఘనిస్తాన్, ఇంగ్లాండ్ మరియు దక్షిణాఫ్రికా చేతిలో ఓడిపోయింది. దూరంగా ఉన్న పర్యటనల్లో పెద్ద మార్జిన్ల తర్వాత.[7] అతను 30 డిసెంబర్ 2014న టెస్ట్ ఫార్మాట్ నుండి రిటైర్మెంట్ ప్రకటించాడు,[8] మరియు 2017లో T20Iని మరియు ODIని కెప్టెన్సీ నుండి వైదొలగాడు

Select Context Language:

Telugu

Enter Question:

ధోని తన తొలి టెస్ట్ ఏ నగరంలో చేశాడు

Select Question Language:

Telugu

Predict Answer

Dhoni played his ODI debut in which city?

Answer: Chittagong

# Results

Context-Telugu  
Question-Telugu



# Streamlit Question Answering App

Enter Context:

महेंद्र सिंह धोनी एक भारतीय पूर्व अंतरराष्ट्रीय क्रिकेटर हैं, जो 2007 से 2017 तक सीमित ओवरों के प्रारूप में और 2008 से 2014 तक टेस्ट क्रिकेट में भारतीय राष्ट्रीय क्रिकेट टीम के कप्तान और सीएसके के वर्तमान कप्तान हैं। उन्होंने भारत को तीन आईसीसी ट्रोफी में जीत दिलाई, जो किसी भी भारतीय कप्तान द्वारा सबसे अधिक है। भारतीय घरेलू क्रिकेट में उन्होंने बिहार और झारखंड क्रिकेट टीम के लिए खेला। वह इंडियन प्रीमियर लीग में चेन्नई सुपर किंग्स (CSK) के कप्तान हैं। उन्होंने आईपीएल लीग के 2010, 2011, 2018 और 2021 संस्करणों में चैंपियनशिप के लिए कप्तानी की। साथ ही उनकी कप्तानी में चेन्नई सुपर किंग्स (CSK) ने दो बार, 2010 और 2014 में प्रीमियर्स लीग टी20 जीता। लंबा। (3) उन्होंने अपना पहला T20 भी एक साल बाद दक्षिण अफ्रीका के खिलाफ खेला। (4) 2007 में, उन्होंने राहुल द्रविड से ODI कप्तानी संभाली और उन्होंने इस वर्ष भारत के T20 कप्तान के रूप में भी प्रवेश किया। (5) 2008 में, उन्हें टेस्ट कप्तान के रूप में चुना गया था। (6) टेस्ट प्रारूप में उनका कप्तानी रिकॉर्ड मिश्रित था, जिसने सफलतापूर्वक भारत को 2008 में न्यूजीलैंड के खिलाफ श्रृंखला जीत और ऑस्ट्रेलिया के खिलाफ बॉर्डर-गावस्कर ट्रॉफी (2010 और 2013 में घरेलू श्रृंखला) में श्रीलंका, ऑस्ट्रेलिया, इंग्लैंड और दक्षिण अफ्रीका से हार का सामना करना पड़ा। अने कड़ीबंस में बड़े अंतर से। (7) उन्होंने 30 दिसंबर 2014 को टेस्ट प्रारूप से अपनी संबन्धित्वि की घोषणा की, [8] और 2017 में T20 और ODI के कप्तान के रूप में पद छोड़ दिया।

Select Context Language:

Hindi

Enter Question:

कौन से ICC ट्रॉफी धोनी जीत चुके हैं?

Select Question Language:

Telugu

Predict Answer

How many ICC trophies has Dhoni won?

Answer: three

# Results

Context-Telugu  
Question-English

# Demo