AQASH

A Question Answering System for Hindi

thelayers

Nukit Tailor

Vamshi Krishna

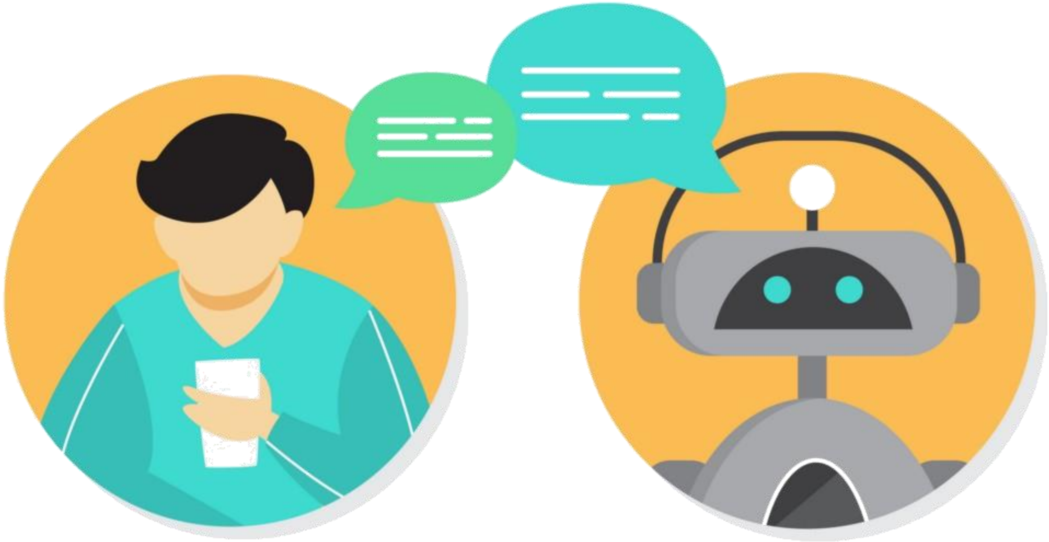
Vanshpreet S. Kohli

Phase - 1

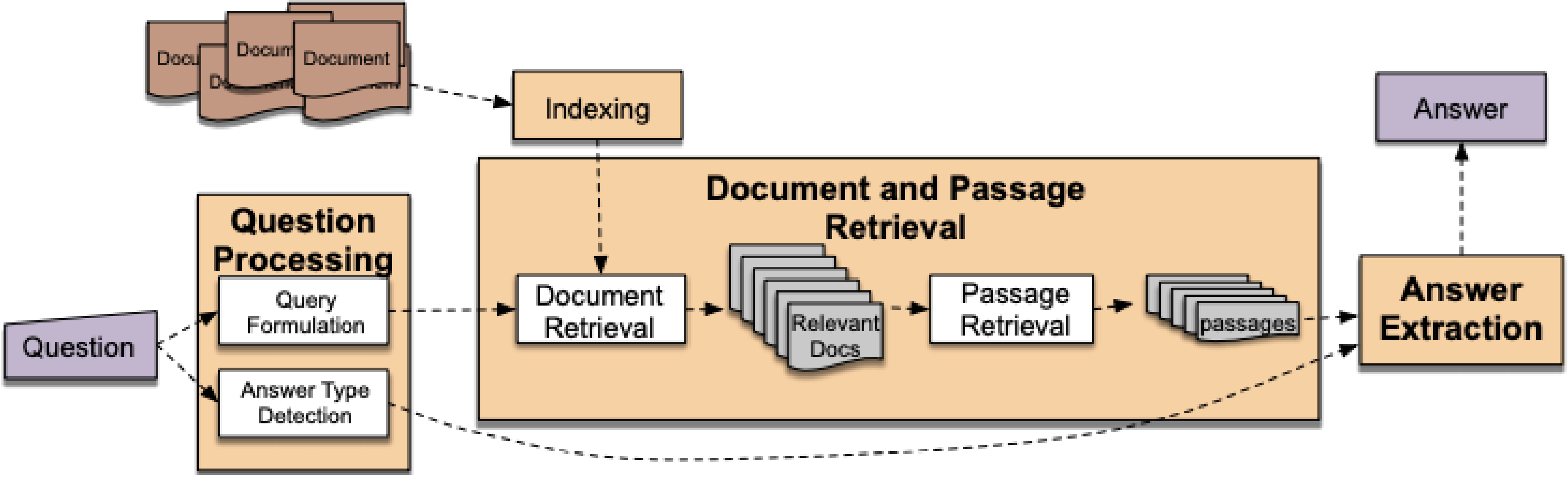
The idea

Indian languages like Hindi, inspite of being spoken by hundreds of millions of people, are underrepresented on the web. One of the most important NLP tools that makes the internet accessible to all is question answering.

Question answering is concerned with building systems that automatically answer questions posed by humans in a natural language. With better question answering systems for Indian languages, we can help Indian users make the most of the web. Predicting answers to questions is a common NLP task for English, but not for Hindi. Here, we attempt to improve baseline models for Q/A systems in Hindi to make the internet more accessible for Hindi speakers.



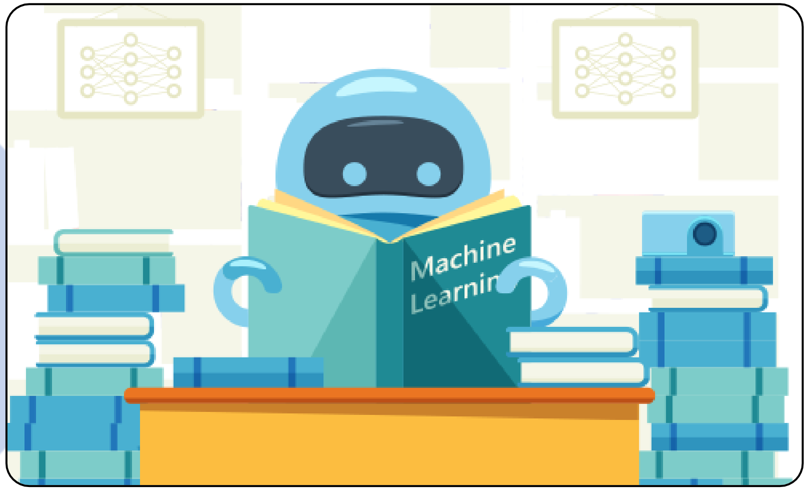
The process


Docu
Docum
Document
Document
Indexing
Answer
Document and Passage Retrieval
Question Processing
Query Formulation
Answer Type Detection
Question
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Answer Extraction

The multilingual models

mBERT

Multilingual BERT (mBERT) was released along with BERT, supporting 104 languages. The approach is very simple: it is essentially just BERT trained on text from Wikipedia content across many languages.



XLM

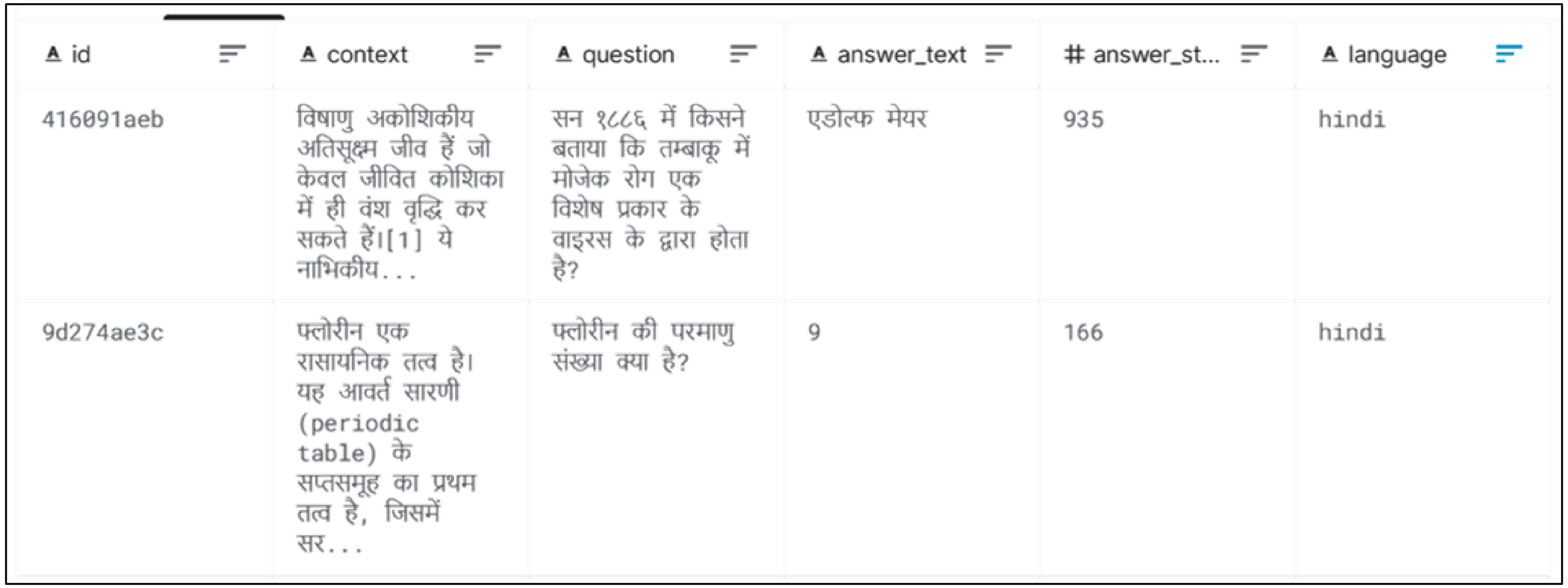
XLM is a Transformer-based model that, like BERT, is trained with the masked language modeling (MLM) objective. Additionally, XLM is trained with a Translation Language Modeling (TLM) objective in an attempt to force the model to learn similar representations for different languages.

XLM-R

XLM-R takes a step back from XLM, and just trains RoBERTa on a huge, multilingual dataset at an enormous scale. Unlabeled text in 100 languages is extracted from CommonCrawl datasets, totaling 2.5TB of text. The only noteworthy difference to RoBERTa is the much larger vocabulary size.

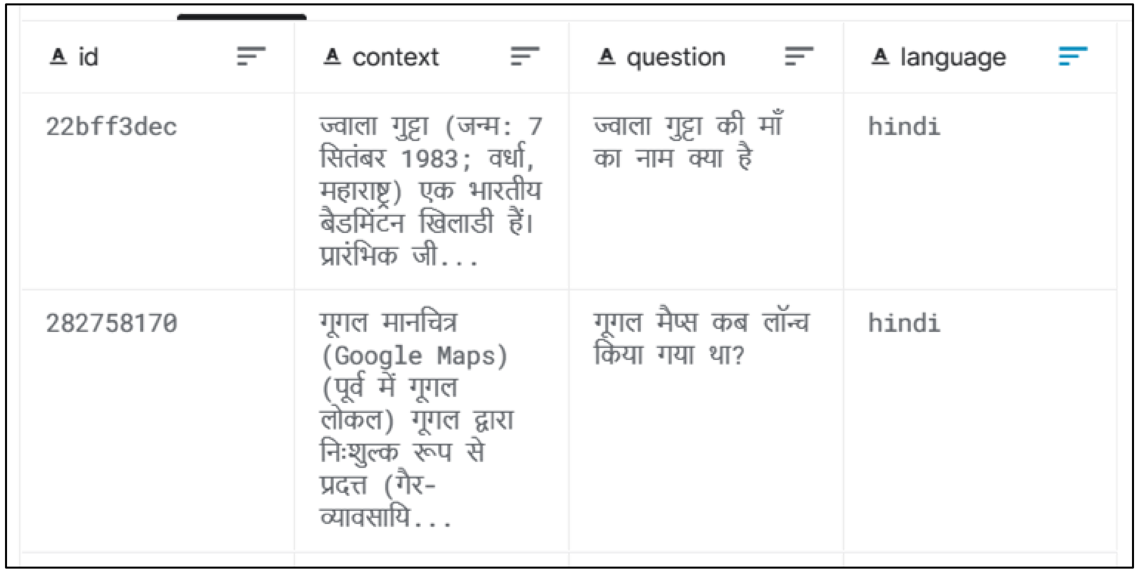
The dataset – chaii 2021

train.csv – 1114 rows



id	context	question	answer_text	answer_st...	language
416091aeb	विषाणु अकोशिकीय अतिसूक्ष्म जीव हैं जो केवल जीवित कोशिका में ही वंश वृद्धि कर सकते हैं।[1] ये नाभिकीय...	सन १८८६ में किसने बताया कि तम्बाकू में मोजेक रोग एक विशेष प्रकार के वाइरस के द्वारा होता है?	एडोल्फ मेयर	935	hindi
9d274ae3c	फ्लोरीन एक रासायनिक तत्व है। यह आवर्त सारणी (periodic table) के सप्तसमूह का प्रथम तत्व है, जिसमें सर...	फ्लोरीन की परमाणु संख्या क्या है?	9	166	hindi

test.csv



id	context	question	language
22bff3dec	ज्वाला गुट्टा (जन्म: 7 सितंबर 1983; वर्धा, महाराष्ट्र) एक भारतीय बैडमिंटन खिलाडी हैं। प्रारंभिक जी...	ज्वाला गुट्टा की माँ का नाम क्या है	hindi
282758170	गूगल मानचित्र (Google Maps) (पूर्व में गूगल लोकल) गूगल द्वारा निःशुल्क रूप से प्रदत्त (गैर-व्यावसायि...	गूगल मैप्स कब लॉन्च किया गया था?	hindi

The deliverables

Phase 2

- Exploring different approaches and ML models, and choosing the right stack.
- Making a rudimentary working Q/A system.

Phase 3

- Improving the existing model further and optimizing it by improving its accuracy.
- Proper documentation of the whole project
- Interface for the project

Phase - 2

Exploring different ML models

XLM

Model fine-tuned
for multilingual data

- Fine-tuned on XQuaD for multilingual Q&A (11 languages, including Hindi)
- Includes ~ 100 languages
- Jaccard score: 0.007 (chaii)
- Link: https://bit.ly/3GCKweO

XLM-R

English base
model

- Trained on SQuAD 2.0 dataset
- Not a multilingual model, but performs better than XLM
- Jaccard score: 0.349 (chaii)
- Link: https://bit.ly/3q00A4p

Exploring different ML models

XLM-R

Base multilingual model

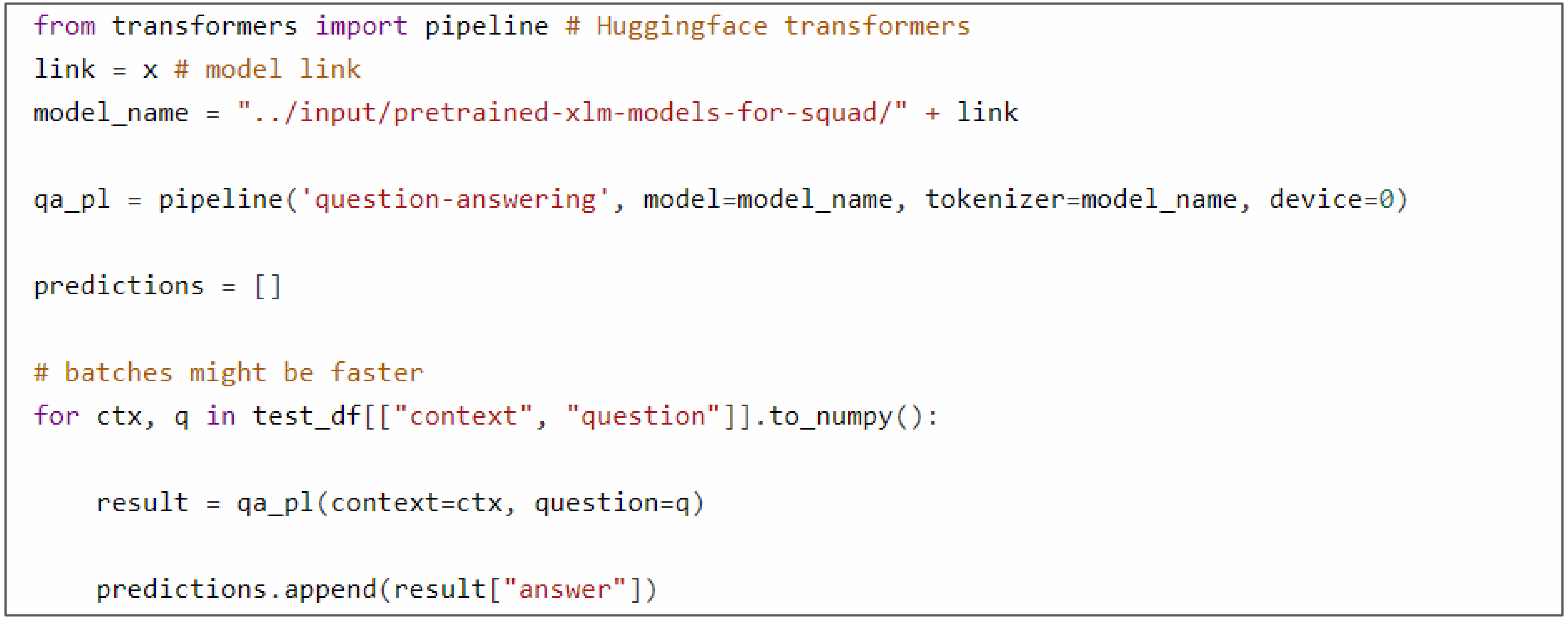
- Trained on SQuAD 2.0 dataset
- Evaluated on German XQuaD and German MLQA
- Jaccard score: 0.493 (chaii)
- Link: https://bit.ly/3GHiezT

XLM-R

Larger multilingual model

- Trained on SQuAD 2.0 dataset on larger hyper-parameters
- Evaluated on German XQuaD and German MLQA
- Jaccard score: 0.571 (chaii)
- Link: https://bit.ly/2ZPgZhm

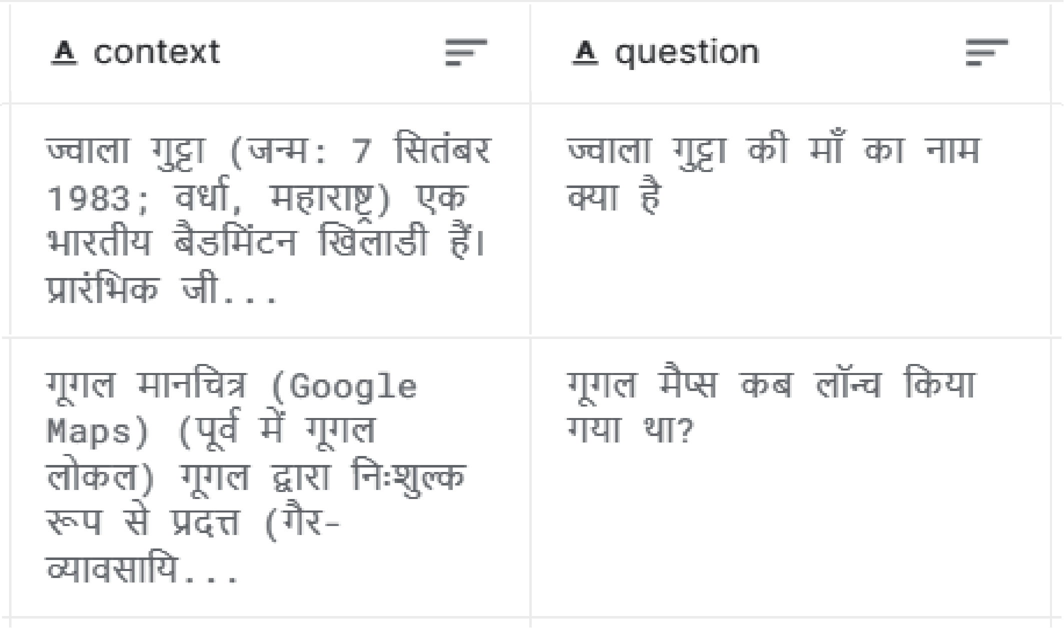
Making a rudimentary model



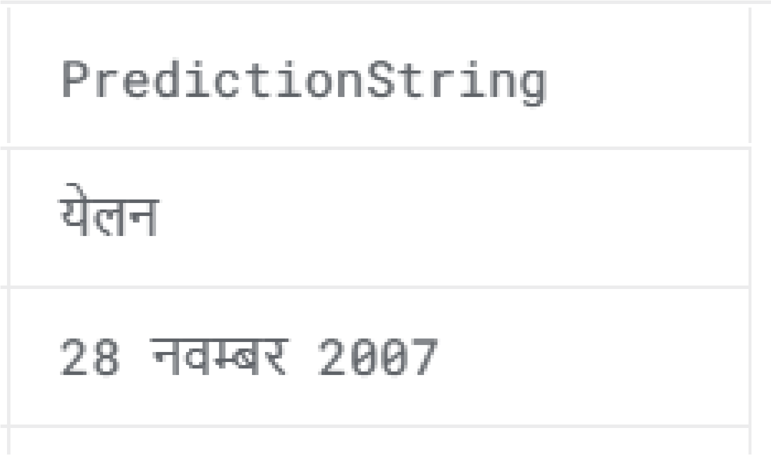
```
from transformers import pipeline # Huggingface transformers
link = x # model link
model_name = "../input/pretrained-xlm-models-for-squad/" + link

qa_pl = pipeline('question-answering', model=model_name, tokenizer=model_name, device=0)

predictions = []

# batches might be faster
for ctx, q in test_df[["context", "question"]].to_numpy():

    result = qa_pl(context=ctx, question=q)

    predictions.append(result["answer"])
```

Using the Huggingface pre-trained model with the pipeline interface; no fine-tuning

The results



context	question
ज्वाला गुट्टा (जन्म: 7 सितंबर 1983; वर्धा, महाराष्ट्र) एक भारतीय बैडमिंटन खिलाडी हैं। प्रारंभिक जी...	ज्वाला गुट्टा की माँ का नाम क्या है
गूगल मानचित्र (Google Maps) (पूर्व में गूगल लोकल) गूगल द्वारा निःशुल्क रूप से प्रदत्त (गैर-व्यावसायि...	गूगल मैप्स कब लॉन्च किया गया था?

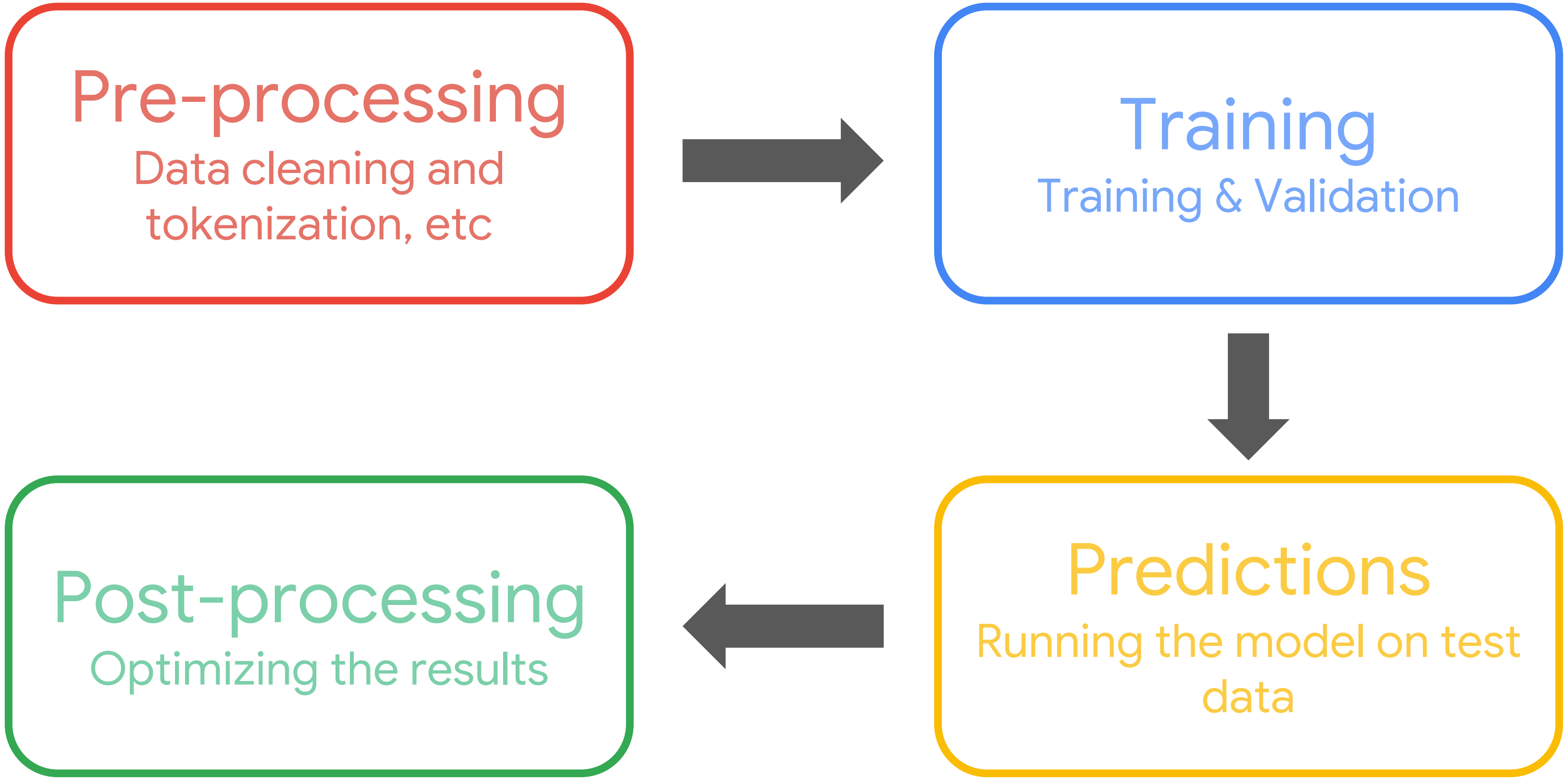


PredictionString
येलन
28 नवम्बर 2007

Future deliverables for phase 3

- Improving the existing model.
- Interface for the project.

The next step: fine-tuning



Pre-processing
Data cleaning and tokenization, etc

Training
Training & Validation

Predictions
Running the model on test data

Post-processing
Optimizing the results

Phase - 3

Overview: phase 3

Fine-tuning the model

Observations and results

Interface

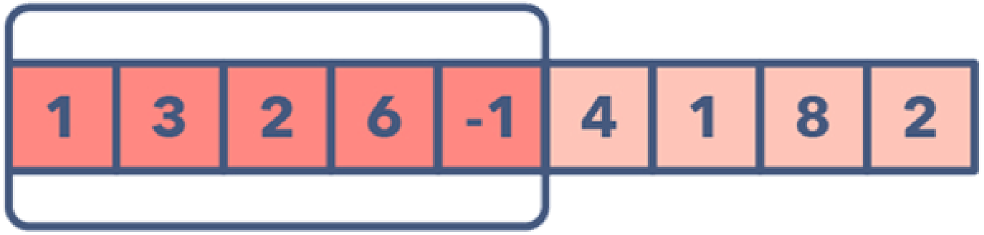
Challenges faced

Fine-tuning the Model

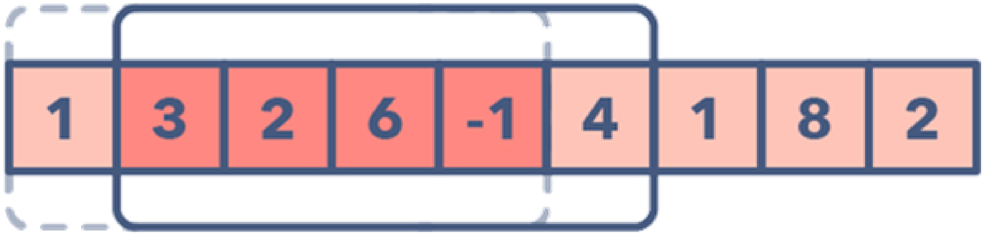
Fine Tuning: Pre-processing

- Tokenization (huggingface)
- Padding and truncation
- The Sliding Window model:

Sliding window -->

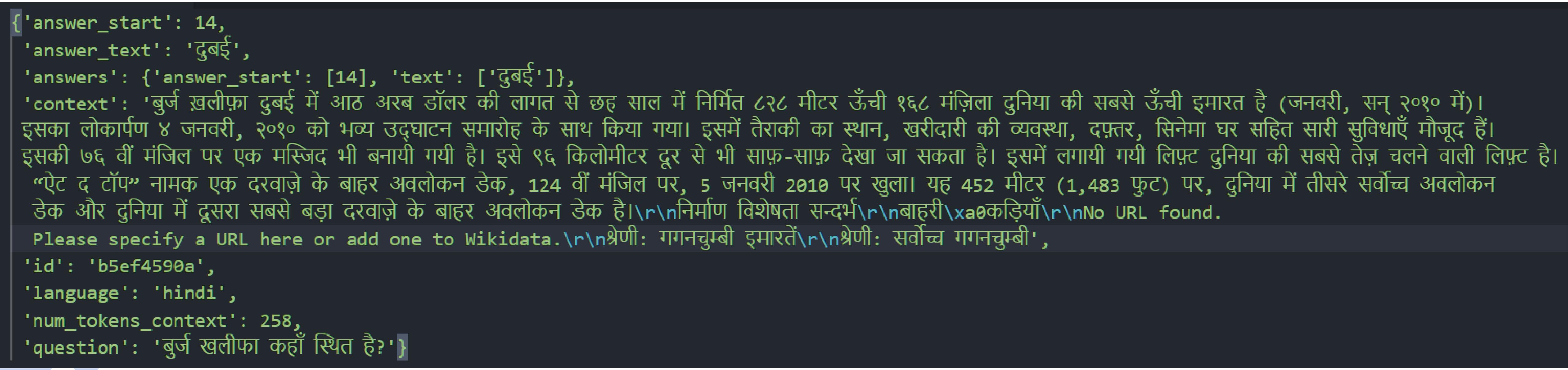


Slide one element forward



Fine Tuning: Training

- Importing the dataset (chaii).
- 700+ examples



```
{'answer_start': 14,
 'answer_text': 'दुबई',
 'answers': {'answer_start': [14], 'text': ['दुबई']},
 'context': 'बुर्ज ख़लीफ़ा दुबई में आठ अरब डॉलर की लागत से छह साल में निर्मित ८२८ मीटर ऊँची १६८ मंज़िला दुनिया की सबसे ऊँची इमारत है (जनवरी, सन् २०१० में)।
इसका लोकार्पण ४ जनवरी, २०१० को भव्य उद्घाटन समारोह के साथ किया गया। इसमें तैराकी का स्थान, खरीदारी की व्यवस्था, दफ़्तर, सिनेमा घर सहित सारी सुविधाएँ मौजूद हैं।
इसकी ७६ वीं मंजिल पर एक मस्जिद भी बनायी गयी है। इसे ९६ किलोमीटर दूर से भी साफ़-साफ़ देखा जा सकता है। इसमें लगायी गयी लिफ़्ट दुनिया की सबसे तेज़ चलने वाली लिफ़्ट है।
 “ऐट द टॉप” नामक एक दरवाज़े के बाहर अवलोकन डेक, 124 वीं मंजिल पर, 5 जनवरी 2010 पर खुला। यह 452 मीटर (1,483 फुट) पर, दुनिया में तीसरे सर्वोच्च अवलोकन
 डेक और दुनिया में दूसरा सबसे बड़ा दरवाज़े के बाहर अवलोकन डेक है।\r\nनिर्माण विशेषता सन्दर्भ\r\nबाहरी\xa0कड़ियाँ\r\nNo URL found.
 Please specify a URL here or add one to Wikidata.\r\nश्रेणी: गगनचुम्बी इमारतें\r\nश्रेणी: सर्वोच्च गगनचुम्बी',
 'id': 'b5ef4590a',
 'language': 'hindi',
 'num_tokens_context': 258,
 'question': 'बुर्ज खलीफा कहाँ स्थित है?'}
```

Fine Tuning: Training

Hyperparameters used (XLMR base)

- Gradient accumulation steps = 8
- Batch size = 4
- Learning rate = 3e – 5
- Number of epochs = 1
- Weight decay = 0.01

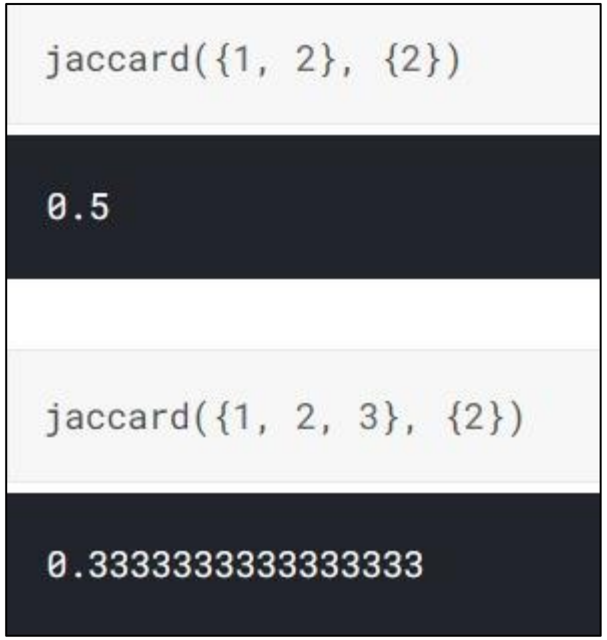
Fine Tuning: Predictions & Post-processing

- Preparing validation dataset
- Obtaining predictions
- Post-processing predictions

Fine Tuning: Evaluation

Jaccard Score

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



```
jaccard({1, 2}, {2})
```

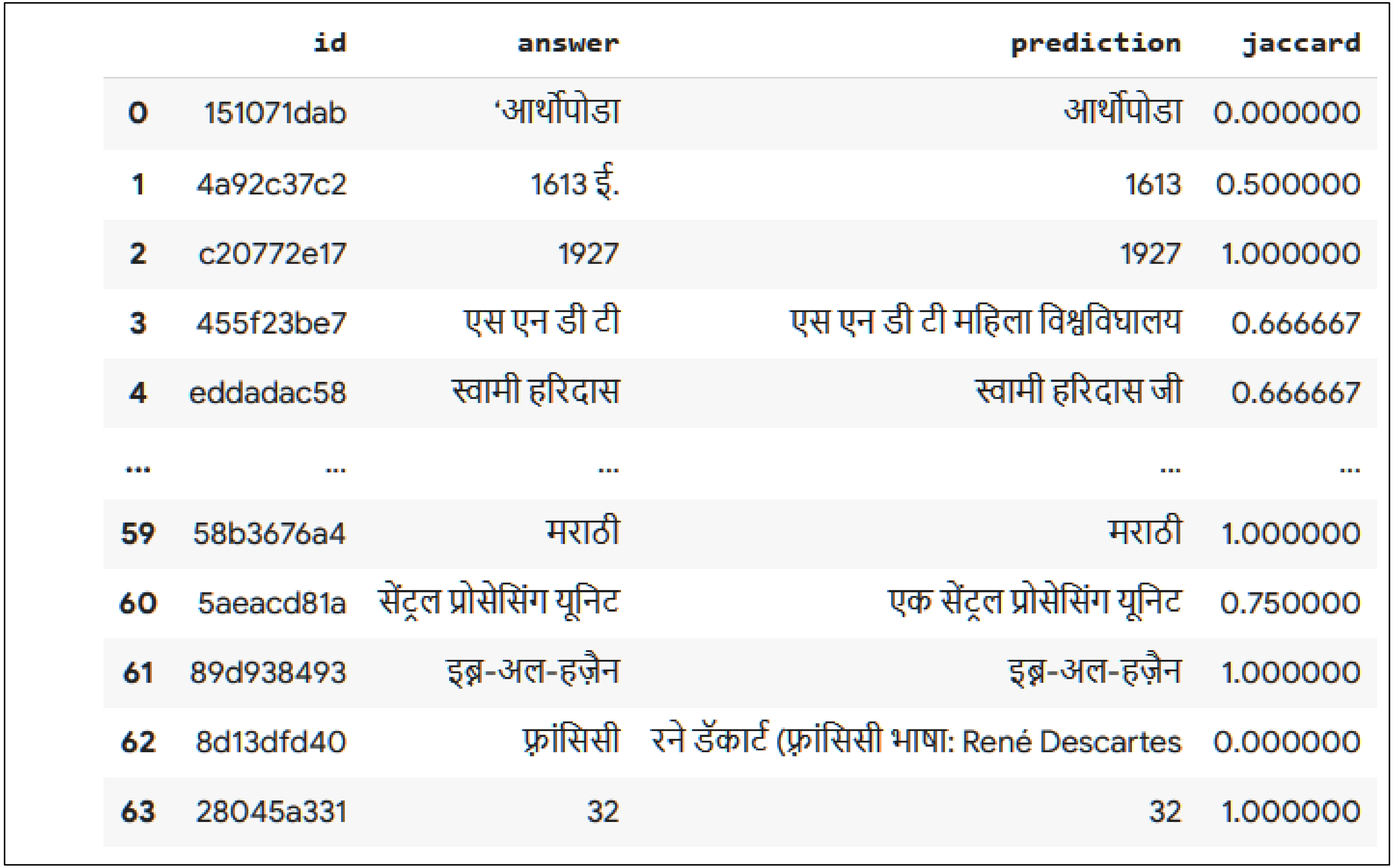
```
0.5
```

```
jaccard({1, 2, 3}, {2})
```

```
0.3333333333333333
```

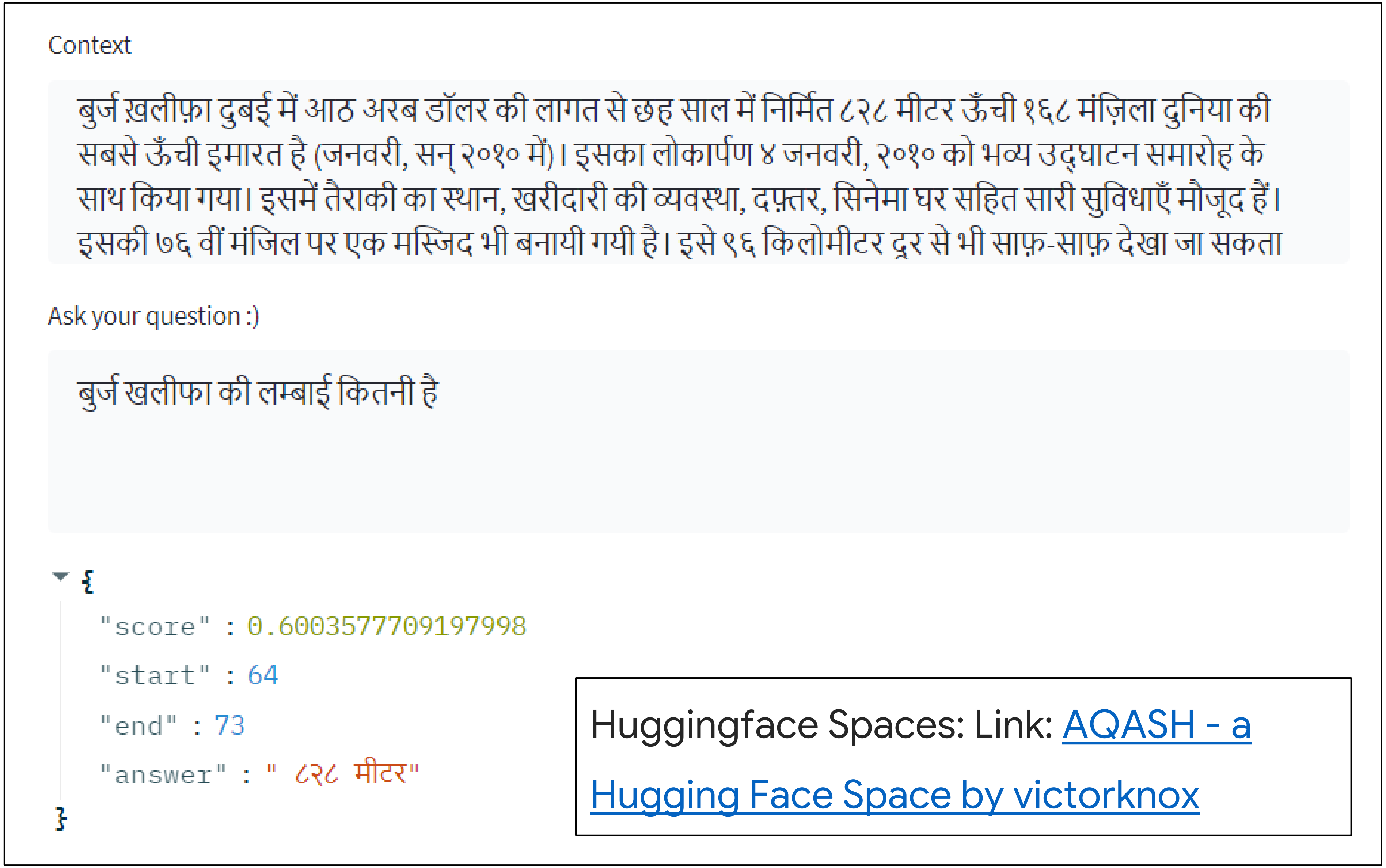
- Obtained mean Jaccard score: 0.77135

Fine Tuning: Evaluation



	id	answer	prediction	jaccard
0	151071dab	‘आर्थोपोडा	आर्थोपोडा	0.000000
1	4a92c37c2	1613 ई.	1613	0.500000
2	c20772e17	1927	1927	1.000000
3	455f23be7	एस एन डी टी	एस एन डी टी महिला विश्वविघालय	0.666667
4	eddadac58	स्वामी हरिदास	स्वामी हरिदास जी	0.666667
...	...	...	...	...
59	58b3676a4	मराठी	मराठी	1.000000
60	5aeacd81a	सेंट्रल प्रोसेसिंग यूनिट	एक सेंट्रल प्रोसेसिंग यूनिट	0.750000
61	89d938493	इब्न-अल-हज़ैन	इब्न-अल-हज़ैन	1.000000
62	8d13dfd40	फ़्रांसिसी	रने डॅकार्ट (फ़्रांसिसी भाषा: René Descartes	0.000000
63	28045a331	32	32	1.000000

Interface



Context

बुर्ज ख़लीफ़ा दुबई में आठ अरब डॉलर की लागत से छह साल में निर्मित ८२८ मीटर ऊँची १६८ मंज़िला दुनिया की सबसे ऊँची इमारत है (जनवरी, सन् २०१० में)। इसका लोकार्पण ४ जनवरी, २०१० को भव्य उद्घाटन समारोह के साथ किया गया। इसमें तैराकी का स्थान, खरीदारी की व्यवस्था, दफ़्तर, सिनेमा घर सहित सारी सुविधाएँ मौजूद हैं। इसकी ७६ वीं मंजिल पर एक मस्जिद भी बनायी गयी है। इसे ९६ किलोमीटर दूर से भी साफ़-साफ़ देखा जा सकता

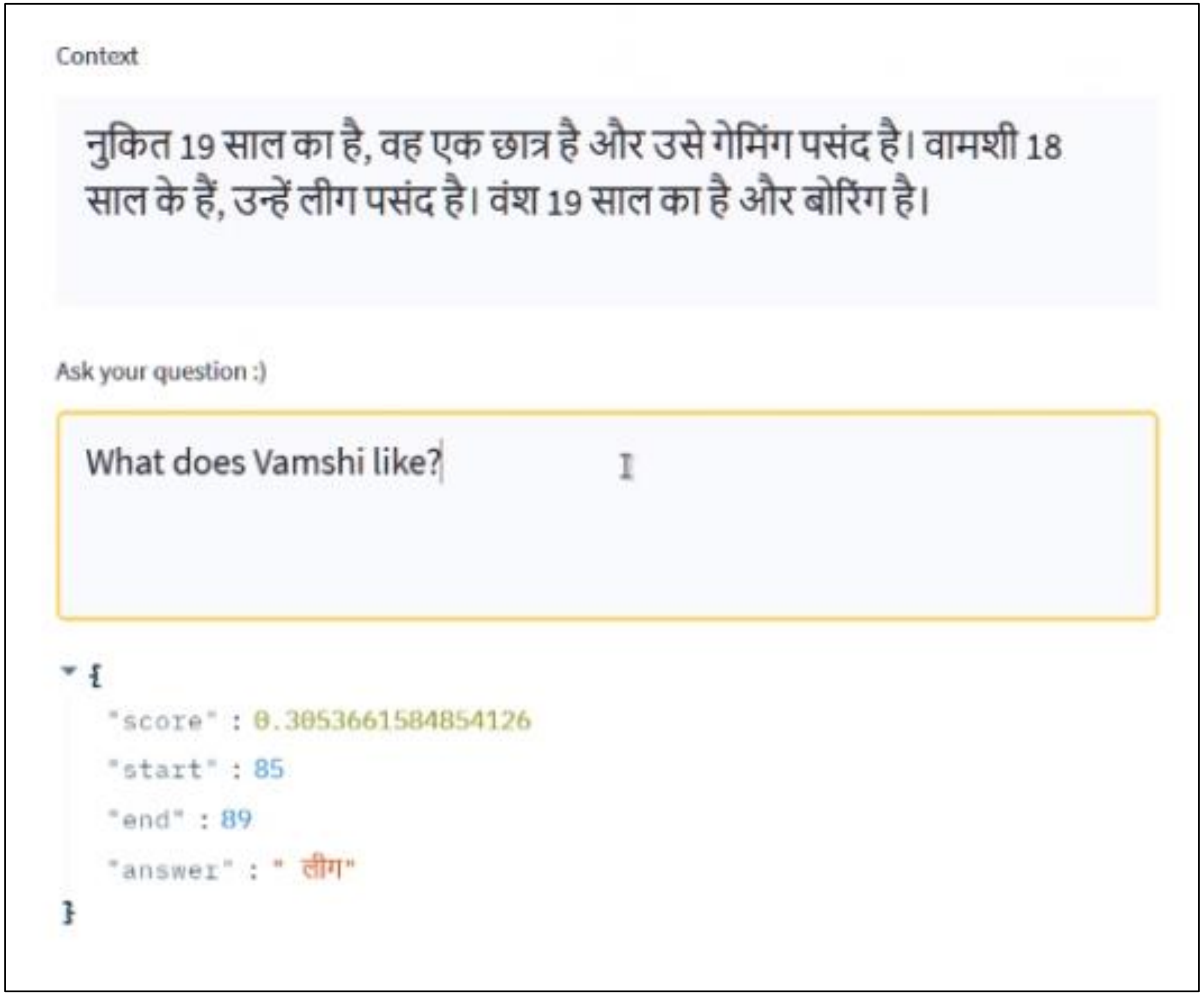
Ask your question :)

बुर्ज खलीफा की लम्बाई कितनी है

```
▾ {
   "score" : 0.6003577709197998
   "start" : 64
   "end" : 73
   "answer" : " ८२८ मीटर"
}
```

Huggingface Spaces: Link: AQASH - a Hugging Face Space by victorknox

Interface



Context

नुकित 19 साल का है, वह एक छात्र है और उसे गेमिंग पसंद है। वामशी 18 साल के हैं, उन्हें लीग पसंद है। वंश 19 साल का है और बोरिंग है।

Ask your question :)

What does Vamshi like?

```
{
  "score" : 0.3053661584854126
  "start" : 85
  "end" : 89
  "answer" : " लीग"
}
```

Challenges faced

- Exploring Question-Answering systems
- Steep learning curve to ML, NLP, Transformers, transfer learning etc as we had no previous knowledge
- Working on Hindi, which is resource-poorer than English
- Training time was huge (~18 hours)

What we learned

- The working of extractive Question-Answering systems and multilingual models
- Application of huggingface transformers and fine-tuning them, etc

References:

A question answering system using machine learning approach – 2016

Prashnottar: a Hindi question answering system – 2012

A Deep Neural Network Framework for English Hindi Question Answering - 2019

Huggingface course

Question-Answering

Chaii

Thank you!