Natural Question Answer Research

Guide : Prof. D V L N Somayajulu
Group : Ashutosh Hathidara(116CS0007)
Nischay Pandey(116CS0021)
Akash Dhayal(116CS0037)

Objective

Main Goal ?

To Improve micro F1-score by modifying existing models.

Build Map

Preprocessing
(Akash & Nischay)

Vectorizing texts
(Akash & Nischay)

Background Research
for Model
(Nischay & Ashutosh)

Tensorflow
implementation
(Ashutosh)

Final Model
(Ashutosh & Nischay)

NQA Dataset
Collection
(Ashutosh)

Ref. Icons GCP Github Repo.
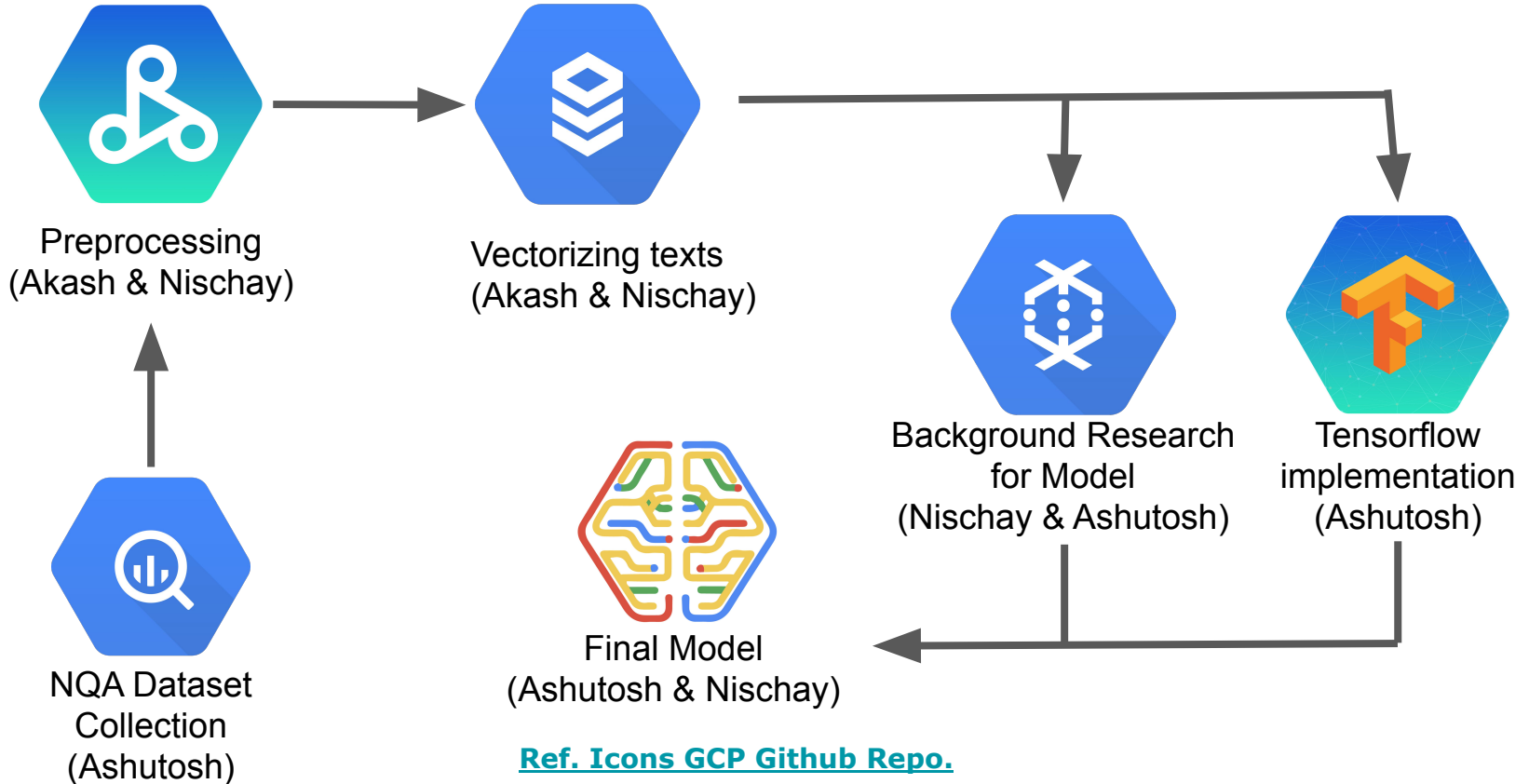
Datasets Used

- [Jian Zhang's SQuAD 2016](#)

- [Google AI's NQA benchmark Dataset 2019](#)

Parameters

- **document_text**
- **question_text**
- **Long_answer_candidates**
- **Annotations**
- **Document_url**
- **example_id**

To Predict

- **long_answer_range** - Range of the long answer in terms of index of words in the document if long answer is found, else it will predict [-1,-1]
- **short_answer_range** - Range of the short answer in terms of index of words in the document if short answer is found, else it will predict [-1,-1]

Evaluation Parameter : F1-Score

Preprocessing & Vectorization

- Position Embedding
- Segment Embedding
- Token embedding

For Example:	This	is	a	boy
	V1	V2	V3	V4

Preprocessing usage : tokenization.FullTokenizer(object)
Embedding Usage : BERT

Ref. Jonathan Bratt's tokenization 2012

Word2Vec (By Tomas Mikolov 2013)
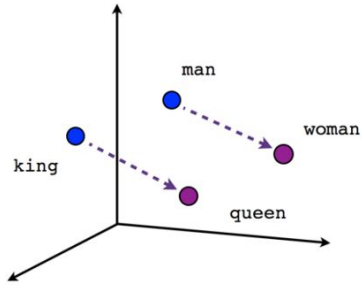
Male-Female
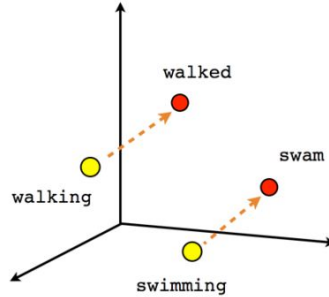
Verb tense
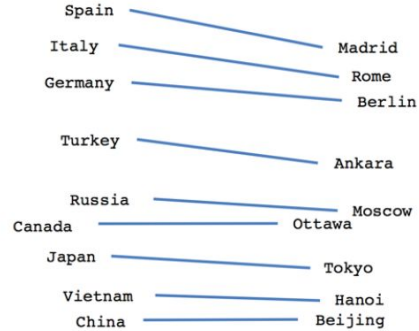
Country-Capital

Drawback :

Context

We went to the river bank.

I need to go to bank to make a deposit.

Context
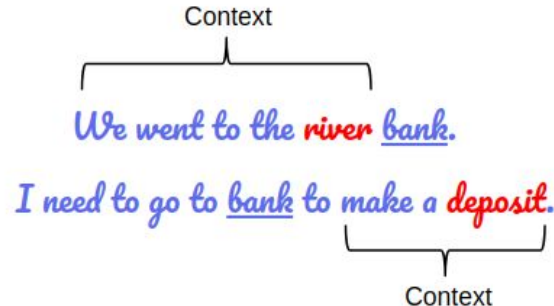
BERT (by Google AI 2019)
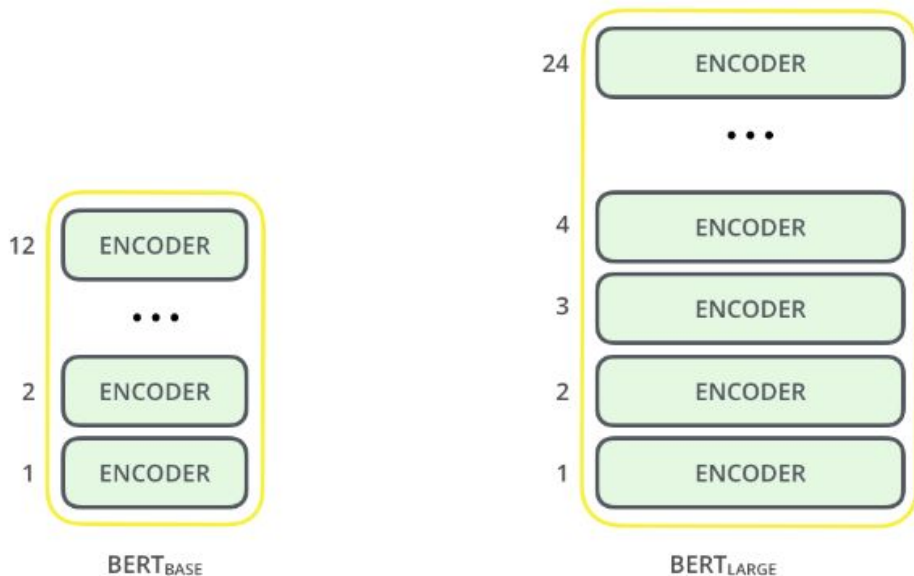
Ref. Blog Analytics Vidya 2019

- BERT Base: 12 layers (transformer blocks), 12 attention heads, and 110 million parameters
- BERT Large: 24 layers (transformer blocks), 16 attention heads and, 340 million parameters

Training on Cloud (on kaggle)

Masking :

Input: The man went to the [MASK]₁ . He bought a [MASK]₂ of milk .
Labels: [MASK]₁ = store; [MASK]₂ = gallon

Ref. Google AI Blog on BERT 2019

Context Recognition:

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

Ref. Google AI Blog on BERT 2019

Implementation

Ref. Tensorflow 2.0

```
{
    "attention_probs_dropout_prob": 0.1,
    "hidden_act": "gelu",
    "hidden_dropout_prob": 0.1,
    "hidden_size": 1024,
    "initializer_range": 0.02,
    "intermediate_size": 4096,
    "max_position_embeddings": 512,
    "num_attention_heads": 16,
    "num_hidden_layers": 24,
    "type_vocab_size": 2,
    "vocab_size": 30522
}
```

Ref. Implementation

Conclusion

F1-score on test : 0.70

SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	87.433	93.160
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

Ref. Google AI Blog on BERT 2019

Implementation Codes

Refer this github repo bit.ly/nqa_iiitk for Project Report, Presentation slides and complete implementation of end-to-end model.

References

- [BERT Github Repo 2019 by Google AI](#)

- [Jian Zhang's SQuAD 2016](#)

- [Google AI's NQA benchmark Dataset 2019](#)

- [Ming-Wei Chang's BERT Paper 2019](#)

- [Google AI Blog on BERT 2019](#)

- [Effective TensorFlow 2.0 Guide](#)