Text to Query Models on Databases

Lushaank K - 20170010082
Vineesh S - 20170010132
Shiva Sai R - 20170010130
Phani Sainath K - 20170010063

Motivation

- Use Text-to-Query model on Databases especially SQL for database programmers.
- Could also be used in Virtual Assistant and in Speech Recognition Applications.
- Could be helpful for Novice Database Programmers and save time for Professional Programmers.

Dataset : WIKISQL

A large crowd-sourced dataset for developing natural language interfaces for relational databases with 80,654 pairs of questions and the corresponding human-verified SQL queries. The massive dataset has attracted much attention in the community and witnessed a significant progress through task-specific end-to-end neural models

Question, query and table ID

These files are contained in the `*.jsonl` files. A line looks like the following:


```
{
   "phase":1,
   "question":"who is the manufacturer for the order year 1998?",
   "sql":{
      "conds":[
         [
            0,
            0,
            "1998"
         ]
      ],
      "sel":1,
      "agg":0
   },
   "table_id":"1-10007452-3"
}
```

The Table:


```
{
        "id": "1-1000181-1",
        "header": ["State/territory", "Text/background colour", "Format", "Current slogan", "Current series",
        "rows": [
                ["Australian Capital Territory", "blue/white", "Yaa·nna", "ACT · CELEBRATION OF A CENTURY 2013
                ["New South Wales", "black/yellow", "aa·nn·aa", "NEW SOUTH WALES", "BX·99·HI", "No slogan on c
                ["New South Wales", "black/white", "aaa·nna", "NSW", "CPX·12A", "Optional white slimline serie
                ["Northern Territory", "ochre/white", "Ca·nn·aa", "NT · OUTBACK AUSTRALIA", "CB·06·ZZ", "New s
                ["Queensland", "maroon/white", "nnn·aaa", "QUEENSLAND · SUNSHINE STATE", "999·TLG", "Slogan em
                ["South Australia", "black/white", "Snnn·aaa", "SOUTH AUSTRALIA", "S000·AZD", "No slogan on cu
                ["Victoria", "blue/white", "aaa·nnn", "VICTORIA - THE PLACE TO BE", "ZZZ·562", "Current series
        ]
}
```

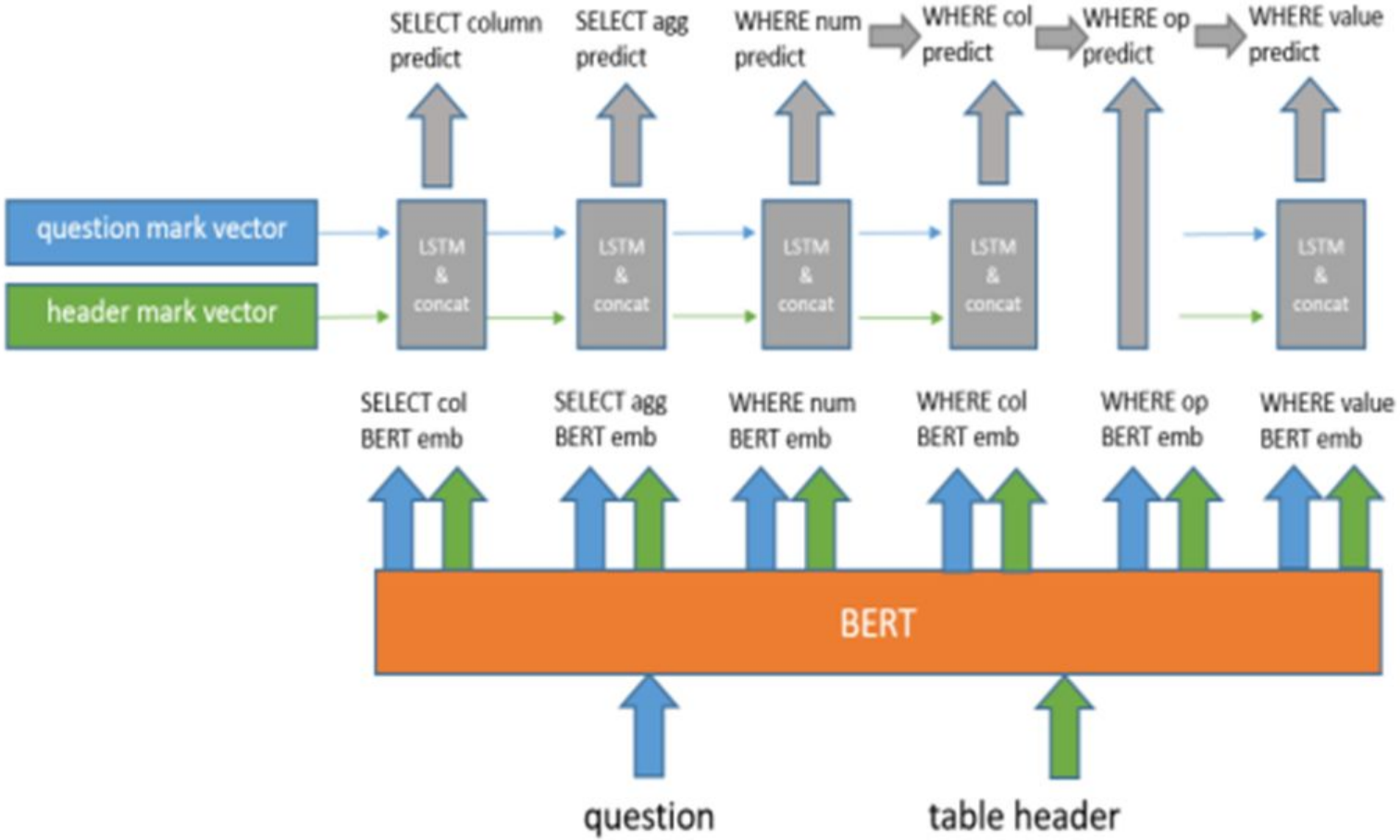
Bert [Bidirectional Encoder Representations from Transformers] Model

Developed by Google and published on October 2018.

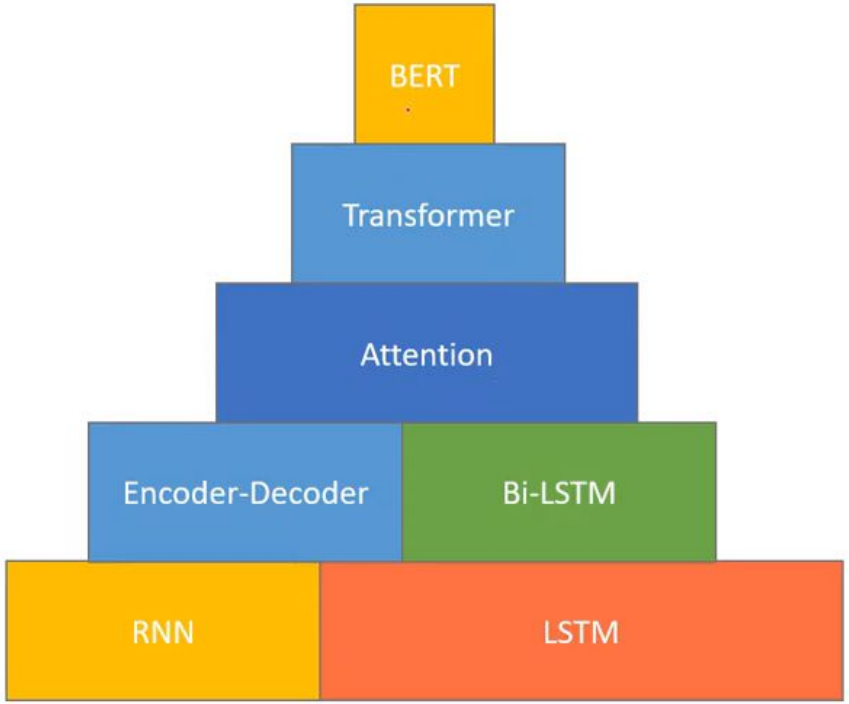
Used in various Natural Language Tasks. It is a large model and expensive to train.

Uses a concept of Transfer Learning. It focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. Eg. Gained knowledge on recognising cars could also be used on recognising trucks.

We use a pre-trained bert model from Google, modify it and performing fine tuning training instead of starting from scratch.


SELECT column predict
SELECT agg predict
WHERE num predict
WHERE col predict
WHERE op predict
WHERE value predict
question mark vector
header mark vector
LSTM & concat
LSTM & concat
LSTM & concat
LSTM & concat
LSTM & concat
SELECT col BERT emb
SELECT agg BERT emb
WHERE num BERT emb
WHERE col BERT emb
WHERE op BERT emb
WHERE value BERT emb
BERT
question
table header

BERT Mountain



Bert Model ...

RNNs - A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs.

LSTM - LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.

Attention - The process in which the program being able to predict the things and items from their priority or from a quick sneak peek from a human is measured here. We try to create a query, key, value pair and to create attention, we determine the relevance between the query and the keys.