Survey Paper on NLIDB Systems
Paper - 1

Survey papers to understand several approaches in this field.

Date 24/09/2018.

In these systems User types a question in english or language to speaker and ML/DL/AI model transforms the question into SQL.

Input *x*:

What is the height of Willis Tower in Chicago?

Rank	Name	Location	Height (ft)	Floor	Year
1	One World Trade Center	New York City	1,776	104	2014
2	Willis Tower	Chicago	1,451	108	1974
...	...	...	...	...	...

Output *y*:

```
SELECT `Height (ft)`
WHERE Name="Willis Tower"
  AND Location="Chicago"
```

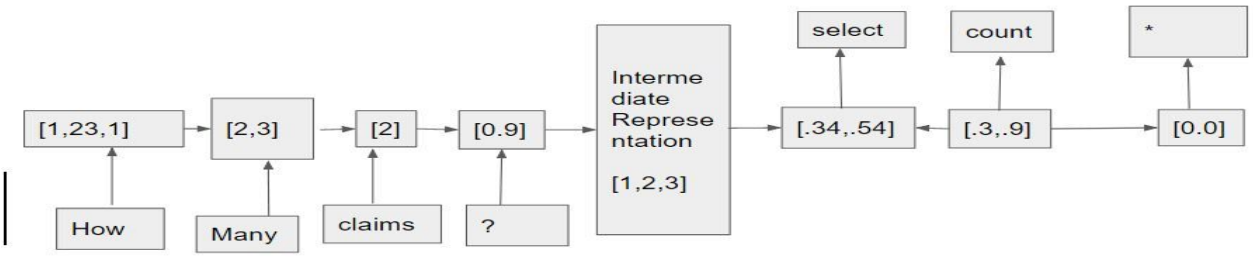
Execution Result:

1,451

Historical Approaches :

<u>Text to Sql</u>

SeqtoSeq Networks:-



This approach is borrowed from current machine translation techniques.Each word in a sentence is encoded[one hot encoding] and embedded sequentially into a i.e intermediate Representation i.e [1,2,3.9] etc. and this intermediate vector is decoded sequentially into sql tokens like select etc.Algorithms like **SqlNet,WikiSql**,**Seq2Sql** are variants of this neural network architecture with “attention component”.

Pros:

Current State of the art accuracies are 87 and improving.

Cons :

Cost of building and maintaining a annotated [english to sql] pairs .
May not generalise well to new database table schemas.
Current approaches are limited to a predefined sql template.
Deep Learning approaches may not yield consistent and reliable results.

Survey Paper on NLIDB Systems
Paper - 1

Survey papers to understand several approaches in this field.

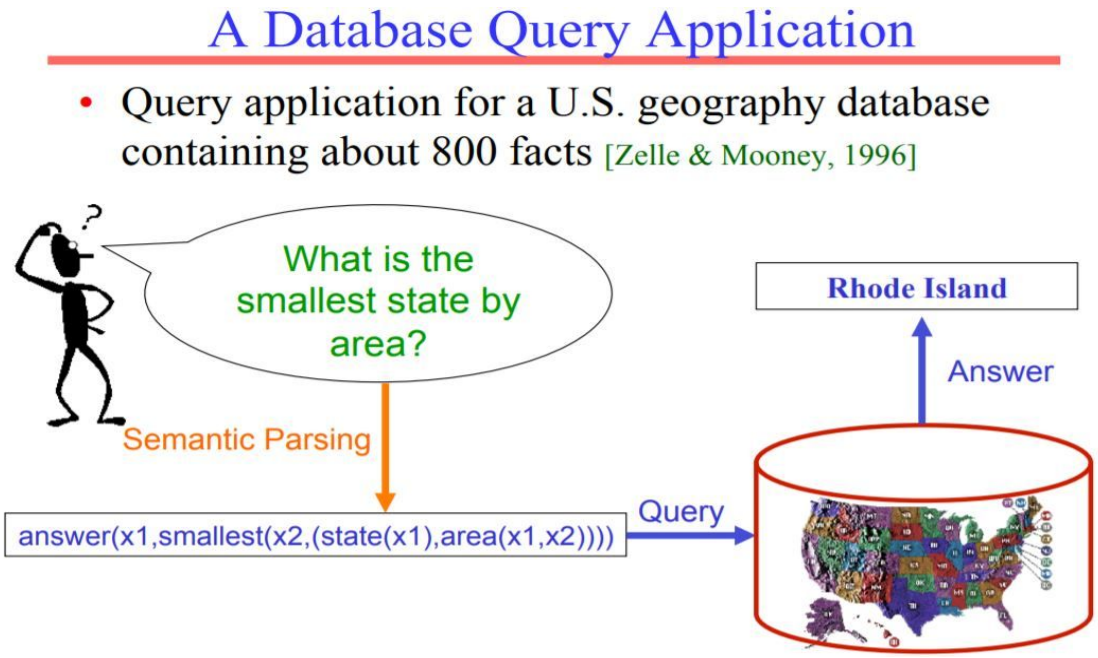
Date 24/09/2018.

Datasets:

Use existing datasets like WikiSql,StackExchange question answer dump ,DBMS Programming assignments from github.
My suggestion is we do not use deep learning approaches for current scenario.These approaches may be more suitable down the line when we have collected .

Executable Semantic Parsing (NLU) (Current Understanding)(Not Yet Complete)

Parsing English sentence into mathematical logic and this logic can be parsed into different target languages like Sql or sparql etc.



Pros :

Can write software in a high level language like java ,python,node js.
Very minimal data requirements.
Can integrate existing libraries to build a pipeline.
Does not require extensive training like Deep Learning .
Semantic parsing is a very mature field.

Cons:

Understand existing source code.

Frameworks that fall into this category i.e Quepy Framework ,Sempre from stanford.

Survey Paper on NLIDB Systems
Paper - 1

Survey papers to understand several approaches in this field.

Date 24/09/2018.

NL2SQL (Current Understanding)(Not yet Complete)

NL2SQL consists of several stages of pipeline.

First Stage:

We construct possible questions that a user gives and tag each and every word of that sentences to either '**SELECT**','**FROM**','**UNK**','**IGN**','**COUNT**'.
Then we take bigram tagger to train on above built corpus.

Currently we do not know the purpose of building such corpus.

Corpus = [('how','select'), ('many','count')]

Second Stage:

First process of pipeline is it to get the all rows from each and every table from database.

We map each value to a column name in a table in the database. [Corpus Building]

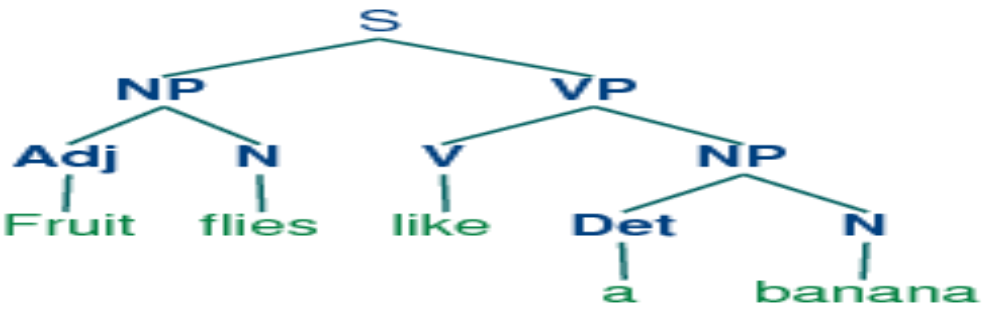
To build feature vectors for a given value in a cell we consider it's left and right values

And their column names and table names.
[v1,v2,v3,v1+c1,v3+c3] - > c2

We train a classifier to take above features to infer correct table name.

Third Stage:

User phrase is parsed into a syntax tree.



Adj = adjective
N = noun

Survey Paper on NLIDB Systems
Paper - 1

Survey papers to understand several approaches in this field.

Date 24/09/2018.

NP = Noun Phrase
V = verb
VP verb phrase
S = Sentence

Fourth Stage.

Every token in the parse tree is classified with the bigram tagger.

* I will explore further stages of nl2sql in next paper.

* I would introduce new techniques in further papers and explore current techniques in depth.