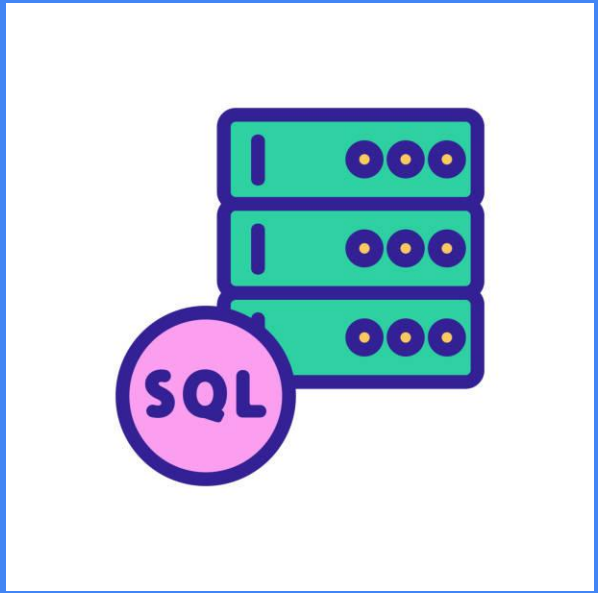
31 March 2022

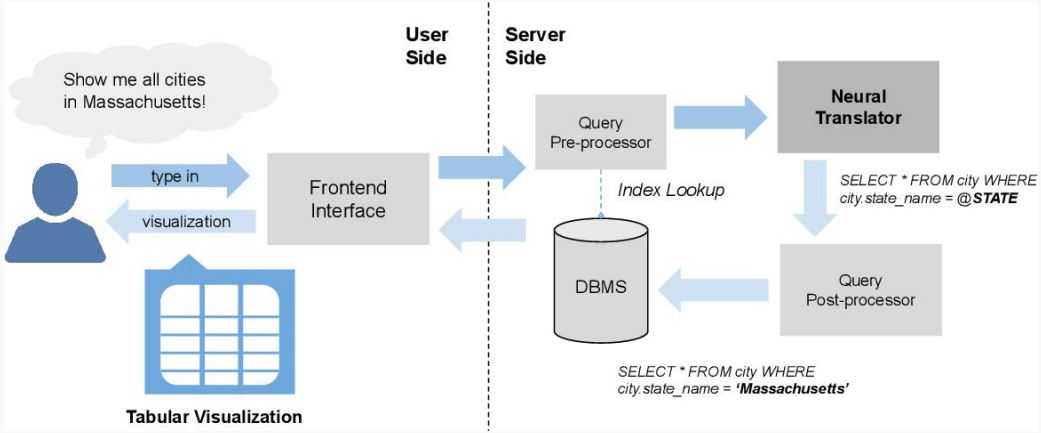
# NL2SQL

Noah Fernandez



# Natural Language to SQL

- Otherwise known as NL2SQL, this expanding field enables the end users without technical skills to interact with relational databases by turning natural language into SQL queries
- A natural language interface database can be interacted with using a statement or question
- NL2SQL unlocks the power of large datasets to users who aren't proficient in query languages



# Early History of NL2SQL

- Several approaches have been made in the field of NL2SQL since it's introduction in the 1960s
- Pattern matching systems: Appeared in the late 1960s and early 1970s, it creates a query from an entry corresponding to a predefined model using a set of rules
- Syntax-based systems: In syntax-based systems, the Natural Language query is parsed syntactically, and the resulting parsing is directly mapped to a Database Query Language expression. These systems use grammatical rules that describe the different possible syntactic structures of queries and a lexicon of words that may appear in the user's queries, LUNAR is the most popular database that is syntax-based.
- The first NLIDBs (natural language interface to databases) had appeared in the late sixties and early seventies and included two very well-known systems:
    - BASEBALL (created to answer questions about baseball games that were played in the American League in that period)
    - LUNAR (answers questions about rock samples brought back from the moon. It managed to answer 80% of the proposed queries without any errors)
- The 60s, 70s, 80s saw many new NLIDBs, but these were designed for specific applications

# Recent History of NL2SQL

- The majority of the existing systems are dependent on the database domain and have been developed for a specific use
- After 1990, new NLIDB approaches were proposed which had a major advantage over previous architectures because they could work independently regardless of the database domain without any need for reconfiguration, meaning they could be used for multiple applications
- Intermediate Representation Languages: Intermediate representation systems were proposed because of the difficulties of translating user queries expressed in Natural Language (NLQ) directly into a database query language. The idea of this method is to first match the NLQ to an intermediate logical query expressed in a representation language. Then, it will be translated into a query in NLIDB. One known system that uses this architecture is PRECISE
- PRECISE is a system developed at the University of Washington. It targets relational databases and the language used to query the database is SQL. The PRECISE system combines linguistic and mathematical approaches to achieve complete independence of information, without any support or configuration

# Current NL2SQL Systems and Capabilities - What they do

- With the rise of deep learning techniques, especially neural networks, the most recent works use a semantic analysis approach which means the normal language words are encoded into semantics and then decoded into SQL queries
- These neural network systems apply the necessary analysis to the input request without taking into consideration it's structure nor the schema of the database, so they are more universal
- Despite active research in text-to-SQL parsing, many contemporary models struggle to develop good representations for a given database schema as well as to properly link column and table references to the question in queries

# Current NL2SQL Systems and Capabilities - How they do it

- Normal language sentences and questions are encoded into semantics and then decoded into SQL queries using a table of potential inputs and then providing corresponding outputs that formulate the responding query
- Neural networks are being used to learn the normal language patterns of end users to generate more accurate query syntax results based on an ever broadening variety of normal language semantics
- Databases that are popular right now are WikiSQL or Spiser use this method

# What’s next for NL2SQL?

- We will most likely see new techniques being employed to improve the NLIDB that are acquired from deep learning performed by the neural networks that are now implemented in these systems
- Neural networks are designed to scan results and begin to recognize patterns that will provide new understanding of the normal language semantics for more accurate syntax to be applied to queries

# NL2SQL Research Opportunities

- More recent NLIDBs such as PHOTON, RatSQL, ValueNet, NLonSpark
- Convolutional and recurrent neural networks

# How would I solve this problem?

- Neural networks are a good start to improving the NL2SQL issues, but the problem that will always remain is despite active research in text-to-SQL parsing, many contemporary models struggle to learn good representations for a given database schema as well as to properly link column and table references in the question. Each database has its own terminology and data is labeled differently for every database, so in order to combat this problem we can use a version of a NLIDB and then a minimal guide is provided to users so instead of learning a lot of technical query language we can simplify it down to keywords to be used in sentences or questions. I think of it as a hybrid between learning how to use SQL and using natural language interfaces to interpret queries.

# References


- https://arxiv.org/pdf/1911.04942.pdf
- https://arxiv.org/pdf/1804.09769.pdf
- http://www.vldb.org/pvldb/vol13/p1737-kim.pdf
- https://arxiv.org/pdf/1911.04942.pdf
- https://arxiv.org/pdf/2108.00804.pdf
- https://2021.ecmlpkdd.org/wp-content/uploads/2021/09/sub_767.pdf
- https://www.e3s-conferences.org/articles/e3sconf/pdf/2021/05/e3sconf_iccsre2021_01039.pdf