Natural Language Querying of Complex Business Intelligence Queries

Jaydeep Sen*, Fatma Özcan*, Abdul Quamar*, Greg Stager+, Ashish Mittal*,

Manasa Jammi*, Chuan Lei*, Diptikalyan Saha*, Karthik Sankaranarayanan*

*IBM Research AI, +IBM Canada

Motivation

- Business Intelligence (BI) queries provide invaluable insights in the enterprise
- NL interfaces enable BI querying for business users, who are not SQL experts, beyond fixed reports
- Existing NLIDB systems fail to handle complex nested SQL queries needed by BI in the enterprise

Overview

- **Extension of our earlier system, ATHENA[1,2]** : A state-of-the-art Ontology Based NLIDB system
- **Ontology is used** to capture the deep domain semantics needed to model the target domain
- Heuristics to detect and guide subquery formations by **combining the use of intelligent lexicon analyzers together with deep domain reasoning over the ontology**
- **Generic and domain agnostic** system and algorithms, capable of generating complex SQL queries involving selections, aggregations, as well as nesting
- Rule-based interpretation, **no need for training data**
- **High accuracy** in preliminary results, proving the effectiveness of using a combination of lexical analyzer and deep domain reasoning

FIBEN: Finance Domain Benchmark Dataset

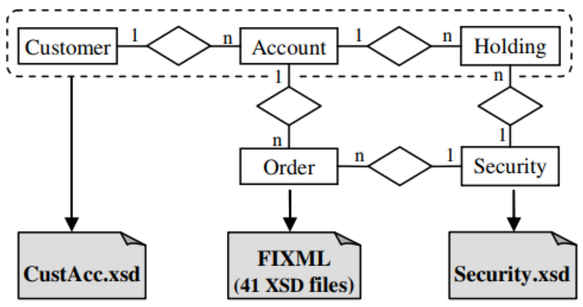
- Emulates real world data mart for a financial application
- Combines SEC data with transactional TPoX[6] data

SEC Data[5]

- Provides information about public companies, their officers and financial metrics
- Dataset extracted from the public SEC filings submitted as XBRL documents
- Data curated by running named entity extraction, and entity resolution by IBM Research

TPoX Data

- Transaction Processing benchmark for financial applications.
- Data generator allows scaling

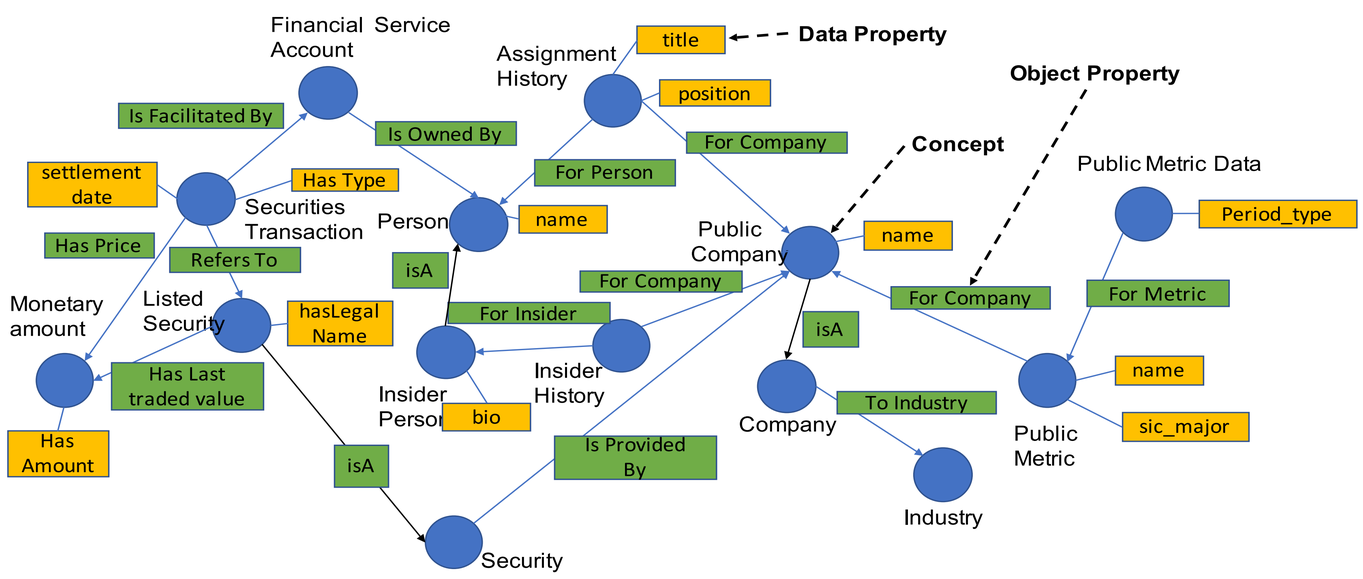


Data transformed to conform to standard finance ontologies:

- FIBO[3] (Finance Industry Business Ontology)
- FRO[4](Finance Report Ontology)

FIBO
Financial Industry Business Ontology

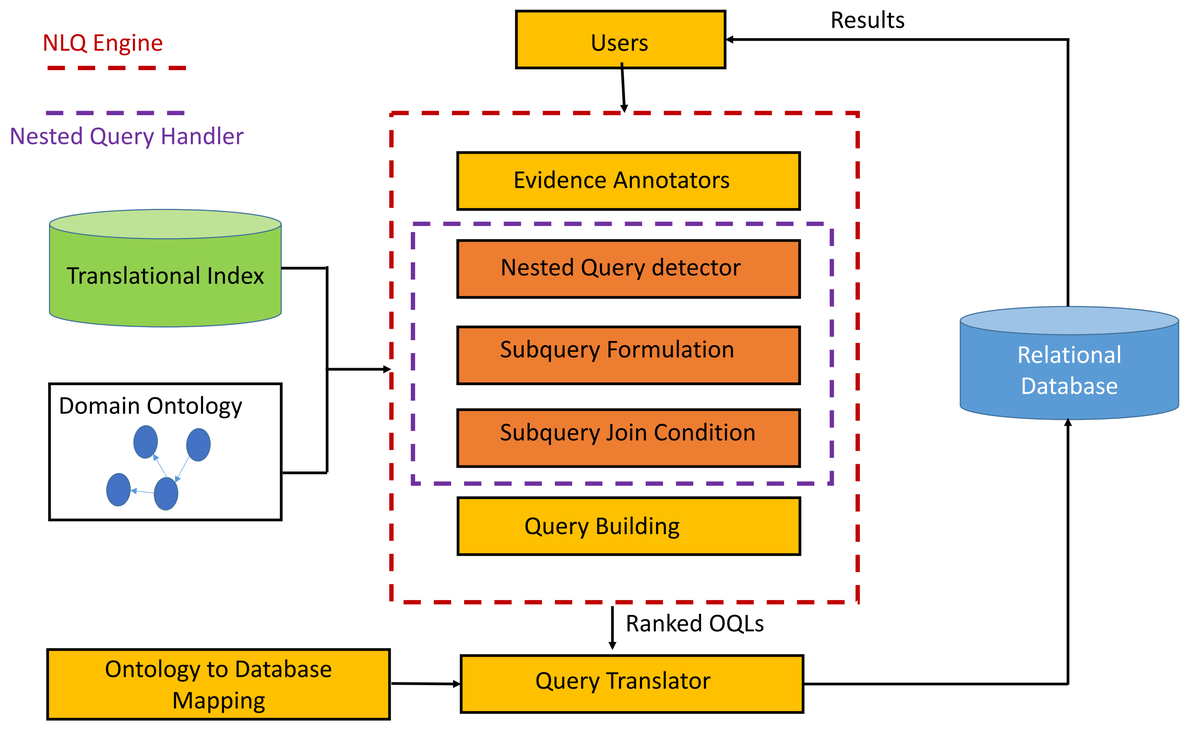
FIBEN Ontology Snapshot



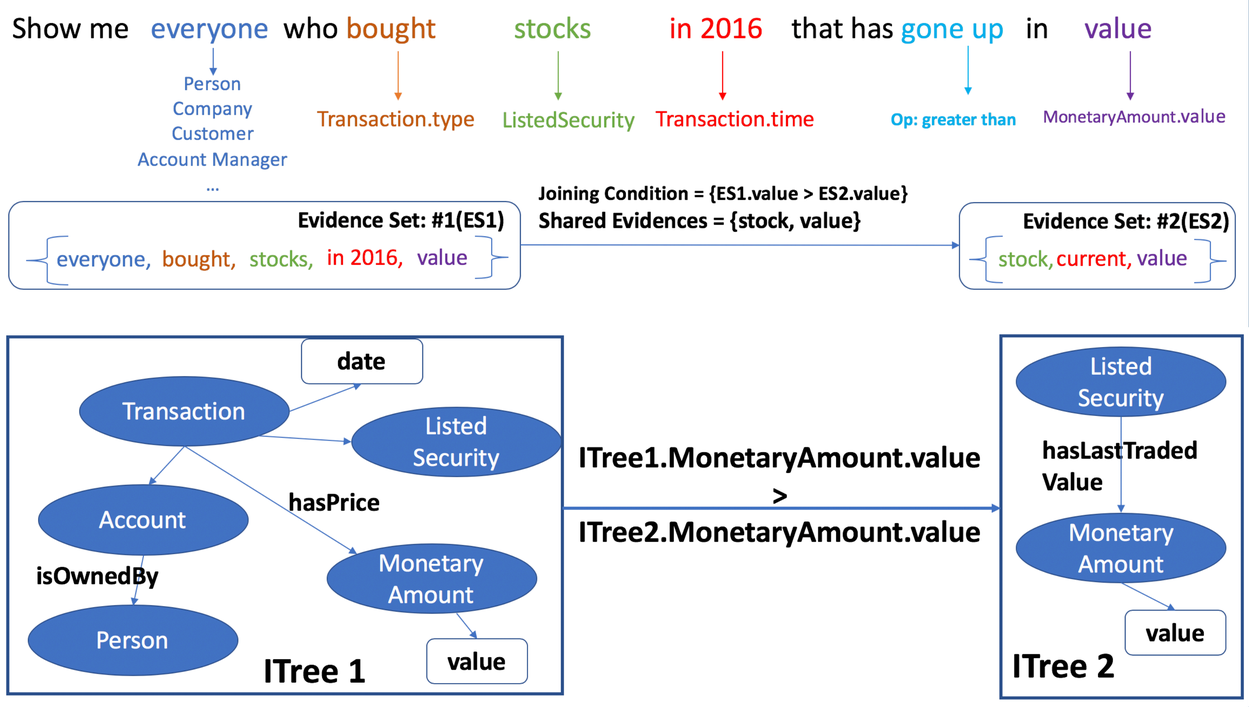
High Level Steps

- **Detection:** Does the input natural language query require nesting?
- **Subquery formation**: If nesting is needed, how to divide the query into subqueries?
- **Subquery Joining**: How to join subquery results to form the complete nested query?

System Architecture



Example Walkthrough



Preliminary Results

Overall Accuracy

Ontology	SQLNest	ATHENA	NALIR	DBPal
FIBEN	92.78	65.35	28.86	41.75

Nested Query Accuracy

Ontology	SQLNest	ATHENA	NALIR	DBPal
FIBEN	79.71	0.0	10.14	21.73

References

1. Diptikalyan Saha,et , "ATHENA: an ontology-driven system for natural language querying over relational data stores", PVLDB 9(12)
2. Shreyas Bharadwaj, et al, "Creation and Interaction with Large-scale Domain-Specific Knowledge Bases", in PVLDB 10(12)
3. FIBO. https://spec.edmcouncil.org/fibo/.
4. FRO. http://xbrl.squarespace.com/financial-report-ontology/
5. SEC Financial Statement Data: https://www.sec.gov/dera/data/financial-statement-data-sets.html.
6. Matthias Nicola, Irina Kogan, and Berni Schiefer, "An XML transaction processing benchmark", in SIGMOD 2007