Presentation on

Searching Using Natural Language Queries

Under Guidance of:
Prof. Animesh Mukherjee
Associate Professor
Prof. Pawan Goyal
Assistant. Professor
Dept. of Comp. Science and Engineering

Mentor:
Suman Kalyan Maity

Presented by:
Group No. 5
Pranjal Kanojiya (16CS60R49)
Rahul Upadhya (16CS60R22)
Samaksh Narayan Garg (16CS60R24)
Sunil Parmar (16CS60R59)
Abhishek Tiwari (16CS60R83)
Mayank Tyagi (16CS60R85)

OBJECTIVES

- Build a search retrieval interface that processes queries typed in natural language.
- Different from traditional Keyword-based search as it deals with meaning of sentence.
- Stemming and removal of stop words is not done.
- Connectors like “how”, “and”, “the” become important.

Example

- Objective : To find the Burj Khalifa Height
- Keyword-based search query: “Burj Khalifa Height”.
- But in Natural Language Query: “How high is the Burj Khalifa?”.

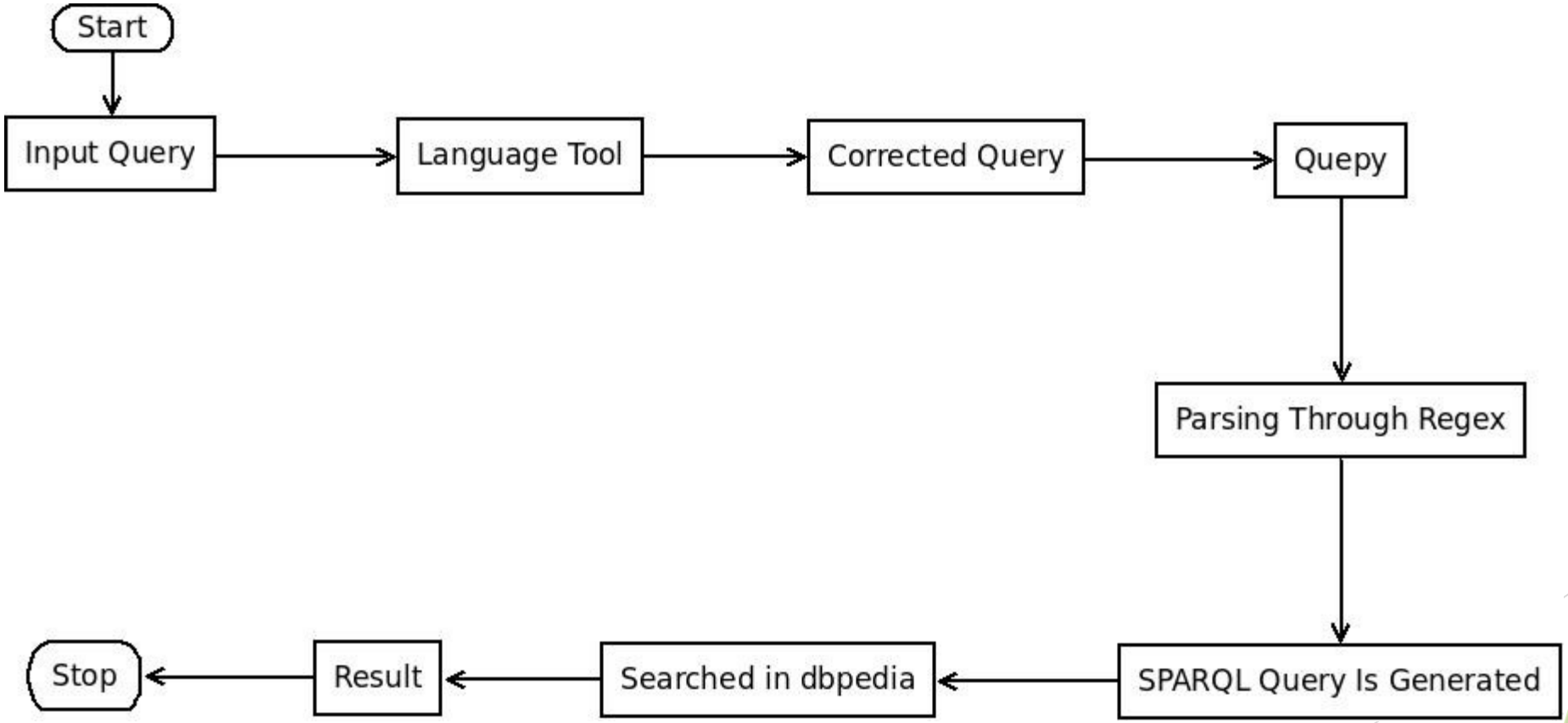
DATASET

- **DBPedia**
  - It is a semantic web that allows users to semantically query relationships and properties of Wikipedia resources.
  - The DBPedia project uses the Resource Description Framework (RDF) to represent the extracted information.
  - Consists of:
    - 3 billion RDF triples,
    - 580 million extracted from the English edition of Wikipedia and
    - 2.46 billion from other language editions

DATASET (contd.)

- For querying it uses RDF(Resource Description Framework) query language
- Data is accessed using an SQL-like query language for RDF called SPARQL

Flow Diagram



APPROACH/IMPLEMENTATION

- Used Language Tool for correcting spelling and grammatical mistakes in the query
- The rules through which the Language Tool makes corrections are specified in grammar.xml
- The corrected query is passed to QUEPY tool in which the query is matched against regular expression that are specified in form of lemmas and POS tags
- In order to write regular expression we used Spacy tool to generate Parse Tree with POS tag.

```
pranjal:irproject$ python parsetree.py
           are_VBP
              |
    __________|____________
   |                   members_NNS
   |                       |
   |          _____________|________________
   |         |                            of_IN
   |         |                              |
Who_WP   the_DT                    Metallica_NNP
```

```
regex1 = Band() + Lemma("member")
regex2 = Lemma("member") + Pos("IN") + Band()
regex3 = Pos("WP") + Lemma("be") + Question(Pos("DT")) + Lemma("member") + \
    Pos("IN") + Band()

regex = (regex1 | regex2 | regex3) + Question(Pos("."))
```

APPROACH/IMPLEMENTATION (contd.)

- If the query matches the regular expression the query is converted to corresponding SPARQL form
- This SPARQL query is then processed on DBPedia to get the results

EXPERIMENTS

- The language tool generates a list of words corresponding to the correction required in the sentence
- We choose the first word proposed by the language tool for replacement in the query
- This word replacement can be improved by considering the context of the word in the query

EXPERIMENTS (contd.)

- For regular expression in QUEPY, REfO is used which matches regular expressions for arbitrary sequences of objects
- No query is generated if the query doesn’t matches the regular expression
- The regular expressions uses lemmas and POS tags to specify similar group of words that are mentioned in WordNet Synsets
- There are various labels through which dataset of DBPedia is accessed and we need to know the database label in which a particular kind of query is processed

Types of Queries Supported

- WHAT IS
- WHO IS
- WHERE IS
- MOVIES(eg. Movies of Brad Pitt)
- Music(eg. Who are the the)
- Person/People(Who is Bill Gates)

DEMO

Search Query Here | Who is Bill Gates ? | Search

Who is Bill Gates ?

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX quepy: <http://www.machinalis.com/quepy#>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX dbpprop: <http://dbpedia.org/property/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x1 WHERE {
  ?x0 rdf:type foaf:Person.
  ?x0 rdfs:label "Bill Gates"@en.
  ?x0 rdfs:comment ?x1.
}
```

William Henry "Bill" Gates III (born October 28, 1955) is an American business magnate, philanthropist, investor, and computer programmer. In 1975, Gates and Paul Allen co-founded Microsoft, which became the world's largest PC software company. During his career at Microsoft, Gates held the positions of chairman, CEO and chief software architect, and was the largest individual shareholder until May 2014. Gates has authored and co-authored several books.

Corrected question

==========

Who is Bill Gates ?

DEMO (contd.)

Search Query Here What is Metalica ? Search

What is Metallica ?

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX quepy: <http://www.machinalis.com/quepy#>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX dbpprop: <http://dbpedia.org/property/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x1 WHERE {
  ?x0 rdfs:label "Metallica"@en.
  ?x0 rdfs:comment ?x1.
}
```

Metallica is an American heavy metal band formed in Los Angeles, California. Metallica was formed in 1981 when vocalist/guitarist James Hetfield responded to an advertisement posted by drummer Lars Ulrich in a local newspaper. The band's current line-up comprises founding members Hetfield and Ulrich, longtime lead guitarist Kirk Hammett and bassist Robert Trujillo. Lead guitarist Dave Mustaine and bassists Ron McGovney, Cliff Burton and Jason Newsted are former members of the band. Metallica collaborated over a long period with producer Bob Rock, who produced all of the band's albums from 1990 to 2003 and served as a temporary bassist between the departure of Newsted and the hiring of Trujillo.

Corrected question
==========
What is Metallica ?

DEMO (contd.)

Search Query Here | Where is Statue of Liberty ? | Search

Where is Statue of Liberty ?

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX quepy: <http://www.machinalis.com/quepy#>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX dbpprop: <http://dbpedia.org/property/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x2 WHERE {
  ?x0 rdfs:label "Statue of Liberty"@en.
  ?x0 dbpedia-owl:location ?x1.
  ?x1 rdfs:label ?x2.
}
```

Liberty Island
United States
Manhattan
New York
New York City

Corrected question

==========

Where is Statue of Liberty ?

Future Work

- Domain of Questions can be increased
- Context based spelling correction
- Other extensive database can be used for better results

References

- http://quepy.readthedocs.io/en/latest/tutorial.html
- https://languagetool.org
- www.nltk.org
- David H.D. Warren, Fernando C.N. Pereira, 1982, An efficient easily adaptable system for interpreting natural language queries, Computational Linguistics archive, 110-122
- Andre Freitas, Edward Curry, Natural Language Queries over Heterogeneous Linked Data Graphs: A Distributional-Compositional Semantics Approach, 2014, Proceedings of the 19th international conference on Intelligent User Interfaces, 279-288