Term Project Report (Part 2)

CS 421 – Natural Language Processing – Spring 2019

University of Illinois at Chicago

Anirudh Nigania (anigan2@uic.edu) Kushagradhi Bhowmik (kbhowm6@uic.edu)

1. Introduction

In this part of the project, we convert natural language queries into appropriate SQL queries. We use our categorization module (from Part 1 of the project) to determine which database (from three databases containing information related to movies, music and geography) to execute the SQL query on. We then convert the database result into a natural language representation. For example, if a Yes-No question yields the result 0 following a database query, we conclude that the answer is No.

2. Implementation

To build the SQL queries compositionally, we first generate a constituency parse tree for the sentence using Stanford CoreNLP implementation available in NLTK. We generated the grammar from a corpus of queries, in the form Node_1 → Sub-node_1 Sub-node_2. We then generalize the unit productions by using the height 2 sub-trees, with the tags providing the antecedent of the rule and the words providing the consequent.

We recursively traverse the parse tree in a depth-first fashion. Upon reaching a node whose only child is an actual word from the sentence, we use the appropriate production rule to generate a part of the SQL query. For example, if the tag is "NNP", the child (say "Loren") becomes the semantic value of the production NNP → Loren. This would then contribute to the WHERE sub-clause of the query we are building. We use Named Entity Recognition to determine that the semantic value in this case is a Person and substitute the corresponding table name into the sub-clause.

The Verb tags contribute to the FROM sub-clause. For example, in "Did Madonna sing Papa Do Not Preach?", the production VB → sing contributes the corresponding tables to use and the way to join them.

The SELECT sub-clause is built using tags like WDT/WRB/WP for the WH-questions and VBD/VBZ for the Yes-No questions. That helps us determine if it is a Yes-No question, or a WH-question, and what type if it is the latter. For example, we use productions such as "WDT → Which" to append the corresponding semantic attachment to the SELECT sub-clause.

We use the word-level adjective tags combine with the following noun tagged word to deduce that the Oscar table needs to be joined to query one of the award categories. We also use the 'Nationality' tags obtained from the NER Tagger to determine the corresponding country-of-birth to query.

The primary difficulties we faced were reconciling the FROM sub-clauses contributed by the verb productions with the noun present in the sentence. For example, let us consider the verb “to win”. Depending upon the context in which it appears, we may need to select some or all of the following tables from the movie database: Oscar, Actor, Person, Director, Movie.

Another issue we had to resolve was in the Yes-No questions with the verb “to be”. In these questions, it is again incumbent upon the noun/noun phrases following the Verb to determine the context of the question. The noun/noun phrases additionally also supply the semantic values for the productions leading to the WHERE sub-clause.

The third difficulty was of the form where the song or movie being asked about is a noun phrase or a simple declarative clause (tagged as Clause Level tag S in the Penn Treebank Tags). We resolved this by merging the subparts initially contributed by the individual words when we encounter the S tag when we arrive a that tag while traversing the parse tree.

3. Sample Output

3.1 Yes-No questions

3.1.1

<Question> Is France in Europe?

<Query> SELECT count(*) From Countries Cn Join CountryContinents CC on CC.CountryID == Cn.ID Join Continents Con on Con.id == CC.ContinentID WHERE Cn.name like "%France%" ;

<Answer>Yes

3.1.2

<Question> Did Bigelow win an oscar for best director?

<Query> SELECT count(*) From movie M Join oscar O on M.id = O.movie_id Join Person P on P.id = O.person_id WHERE P.name like "%Bigelow%" and O.type like "%director%" and O.type like "%director%" ;

<Answer>Yes

3.1.3

<Question> Did Madonna sing Papa Do Not Preach?

<Query> SELECT count(*) From Album Al Join Track T on Al.albumID == T.albumID Join Artist P on P.id == Al.artsitID WHERE P.name like "%Madonna%" and T.name like "%Papa Do Not Preach%" ;

<Answer>Yes

3.2 WH-questions

3.2.1

<Question> Where is Rome?

<Query> SELECT Cn.name From Cities C Join Capitals Cap on C.id == Cap.CityID Join Countries Cn on Cn.id == Cap.CountryID WHERE C.name like "%Rome%" ;

<Answer>Italy

3.2.2

<Question> Which Italian actress won the oscar in 1962?

<Query> SELECT P.name From movie M Join oscar O on M.id = O.movie_id Join Person P on P.id = O.person_id WHERE P.pob like "%ITALY%" and O.type like "%best-actress%" and O.year like "%1962%" ;

<Answer>Sophia Loren

3.2.3

<Question> Which album released in 2003?

<Query> SELECT Al.name From Album Al WHERE Al.releaseDate like "%2003%" ;

<Answer>Dangerously in Love