Computation of User- based Query using Natural Language Processing

Priyanka Khurana
Assistant Professor
Information technology
HCTM technical campus

Mittar Vishav
Assistant Professor
Computer science and engg.
HCTM technical campus

Ruchika Yadav
Assistant Professor
Computer science and engg.
HCTM technical campus

ABSTRACT

Everybody find ease and comfort by using their natural languages to communicate. There is large quantity of natural language material and for this reason there is the need of computer to involve in this process. Beside this people need to communicate with machines and people find natural languages natural. Although there are people who thinks, that only people can effectively use natural languages and thus it is inappropriate to bring computers into this arena, there is already evidence, that programs that manipulate language in various ways can be useful. In this paper, we are extracting information for user based natural language query on the health domain. Information retrieval through such Q/A systems is important sources to help physicians make decisions in patient treatment and as a result, to enhance the quality of patient care by retrieving a vast amount of information in response to a specific user query.

Keywords

Algorithm design for input statement understanding module, Results on developer side.

1. INTRODUCTION

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages or we can say natural language processing is a very attractive method of human–computer interaction. Due to the communication gap between the computer and a human, the problem faced by users of most

information retrieval systems is lack of contextual and knowledge based search in present systems for information retrieval. Mostly, in systems, simple keyword matching occurs for user query and results are not relevant to asked question. User in constrained to write query in particular format and cannot communicate in free- form manner. Our goal is to accept completely free-form input query based on health domain, and to serve as knowledge engine that generates powerful results and presents them with maximum clarity. This paper consists of four folds: To check the syntactic structure of the user query by identifying the tags of grammar, Write rules to cover different syntactic structure for English query, by clustering, Map those computational representations to the query, Query the databases for information extraction, to provide semantic and efficient results on health domain.

2. Algorithm design for input statement understanding Module

2.1 Algorithm

Step1: Input statement (query);

Step2: Lexical - Tokenize the statement

Step 3: Tag the tokens

Step 4 : Syntactic – Check for the correct formation of sentence and remove unwanted

tagged words eg: DT, ",", "."

Step 5: Check for separators eg. Conjuctions, split the sentence based on conditions.

Step 6: For each splitted sentence check for input words and output words as follows:

Separate nouns and English 'word'.

If words doesn't exists

i) Input word with all parameters, forming rules for interpretation.

ii) Fill all tables with related to this word.

If words exists with all data

i) If more than one selection possible: For now select manually (during development phase)

Later when sufficient data are available, selection will be through probability associated. If words exists with insufficient data Update all word table associated with the corresponding word.

Step 7: Using database table decide nouns being a input variable in the statement or the output variable.

Query: Who are the Employee living in India?

Nouns:-

Employee, India

Decision words:-

In

Out of two variables one is output and another is input variable.

Employee – outword

India – input word

Step 8: Map the input / output words to XML

For Now Table Decision Words

Words	Description Input Word	Description Output Word
IN	RI	
Of	RI	LO

After concatenation table

Words	Basic form
Living	Live
Stay	Stay

Synonyms Table

Words	Context
Live	Location
Stay	Location

Step 9: Map the XML to Sql query

Saving previous result

Query	Pattern	Sql Query
Who are the employees living in India?	Who/WP are/VBP the/DT Employee/NN living/VBG In/IN India/NNP? /.	Select employee from Place tab where location = 'india'

3. Results On Developer Side

The results of our implementation on developer side is as follows –

1. Tokens formulation - To make a sentence able to be processed by the computer, it is necessary to divide it in chunks or tokens to understand its meaning and structure. If a category is of no meaning for further operation, that category is ignored for further processing.

E.g. - What are the symptoms of Cancer?

It is tokenized using blanks as – What/are/the/symptoms/of/Cancer/?

2. Tagging/Parsing – The tokenized words are then tagged based on their positions and parts of

speech they are acting as in the sentence and are parsed.

E.g. - What/WP are/VBP symptoms/NNS of/IN Cancer/NN?/.

3. Syntactic Markers – It is based on formal semantics and deals a natural language query semantically. It is necessary to define and ignore the syntactic markers for further processing as they do not have semantic contribution in tracking a query semantically. The words which include, "es ","s "are formatted to their basic forms.

E.g. symptoms is changed to symptom and? is removed which does not contribute for semantics.

Done as – What/are/symptom/of/Cancer

4. Construction of Dictionary/Self Learning – All the words are then asked by developer if to save them to construct dictionary and "synonyms" table is updated if unknown words are there.

5. WH-Rule Formation – Developer is asked for saving the rules of the relational words and

interpreting the words positions and relation among them, to process later in the same way for

similar type of pattern query if user asks. The updating is done in "pattern" table.

6. Input /Output Words – As per the rules formed input / output words are evaluated and saved in separate form if two or more simple sentences are there.

7. Semantics – the input words are then checked against synonyms and replaced to common words via mapping.

8. Split Queries – Compound statement queries are splitted and separated, made independent and written to file.

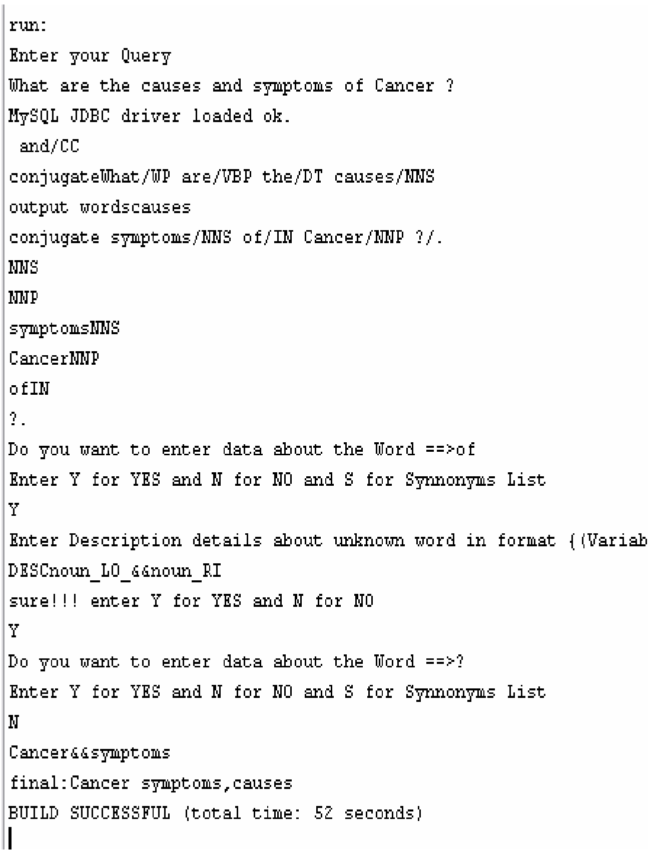
9. XML Generation – the output and input words of each query is read from file are passed to create XML for query structure.

10. SQL Mapping – XML is then mapped to SQL after generation.

11. Query sent to map to database – the query executed in database.

We have crawled the website and stored the result in our database "healthdb "having the data about health domain with attributes – title definition cause symptom diagnosis prevention treatment.

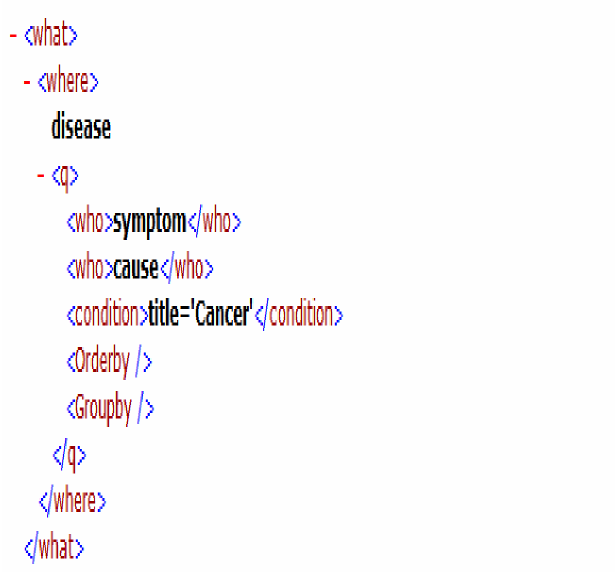
12. Query Executed and results are displayed.



```
run:
Enter your Query
What are the causes and symptoms of Cancer ?
MySQL JDBC driver loaded ok.
 and/CC
conjugateWhat/WP are/VBP the/DT causes/NNS
output wordscauses
conjugate symptoms/NNS of/IN Cancer/NNP ?/.
NNS
NNP
symptomsNNS
CancerNNP
ofIN
?.
Do you want to enter data about the Word ==>of
Enter Y for YES and N for NO and S for Synnonyms List
Y
Enter Description details about unknown word in format {(Variab
DESCnoun_LO_&&noun_RI
sure!!! enter Y for YES and N for NO
Y
Do you want to enter data about the Word ==>?
Enter Y for YES and N for NO and S for Synnonyms List
N
Cancer&&symptoms
final:Cancer symptoms,causes
BUILD SUCCESSFUL (total time: 52 seconds)
```

Figure: 1

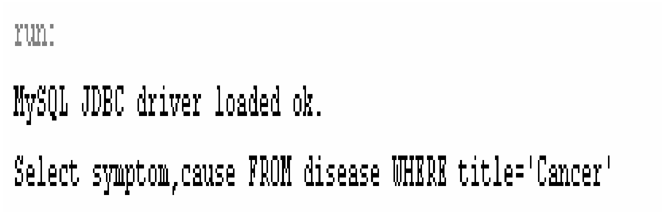
XML Generation:



```
- <what>
  - <where>
      disease
    - <q>
        <who>symptom</who>
        <who>cause</who>
        <condition>title='Cancer'</condition>
        <Orderby />
        <Groupby />
      </q>
    </where>
  </what>
```

Figure: 2

SQl Generation:

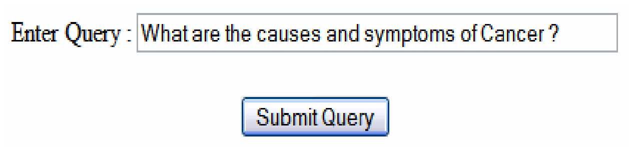


```
run:
MySQL JDBC driver loaded ok.
Select symptom,cause FROM disease WHERE title='Cancer'
```

Figure: 3

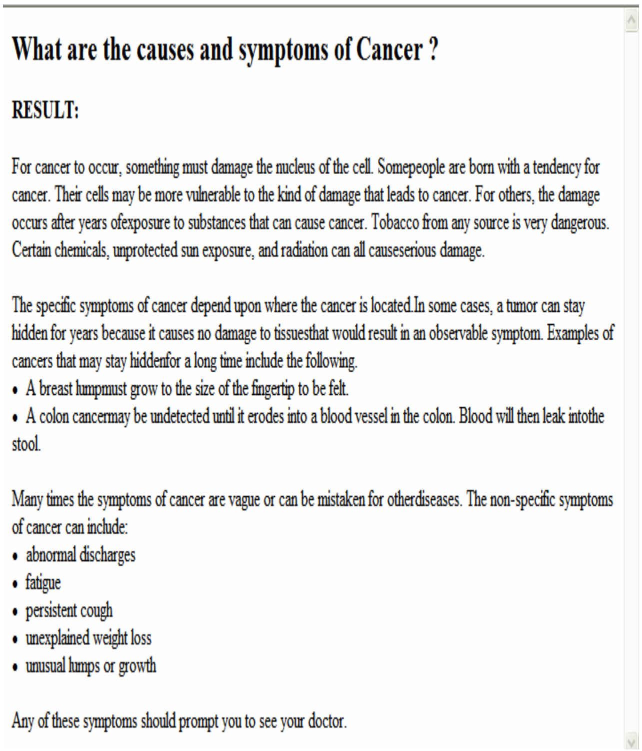
The query is then mapped to synonyms table and results displayed to user

User –Side:



Enter Query : What are the causes and symptoms of Cancer ?

Figure: 4



What are the causes and symptoms of Cancer ?

RESULT:

For cancer to occur, something must damage the nucleus of the cell. Somepeople are born with a tendency for cancer. Their cells may be more vulnerable to the kind of damage that leads to cancer. For others, the damage occurs after years ofexposure to substances that can cause cancer. Tobacco from any source is very dangerous. Certain chemicals, unprotected sun exposure, and radiation can all causeserious damage.

The specific symptoms of cancer depend upon where the cancer is located.In some cases, a tumor can stay hidden for years because it causes no damage to tissuesthat would result in an observable symptom. Examples of cancers that may stay hiddenfor a long time include the following.

- A breast lumpmust grow to the size of the fingertip to be felt.
- A colon cancermay be undetected until it erodes into a blood vessel in the colon. Blood will then leak intothe stool.

Many times the symptoms of cancer are vague or can be mistaken for otherdiseases. The non-specific symptoms of cancer can include:

- abnormal discharges
- fatigue
- persistent cough
- unexplained weight loss
- unusual lumps or growth

Any of these symptoms should prompt you to see your doctor.

Figure: 5

4. Conclusion

The objective is to retrieve the results of free-form input queries asked on health domain doing semantic search by natural language processing (NLP), instead of keyword –matching. This paper consists of four folds: We have checked the syntactic structure of the user query by identifying the tags of grammar, written the rules to cover different syntactic structure for English query by clustering, mapped those computational representations to the query and queried the databases for information extraction, to provide semantic and efficient results on health domain. As per now we are done with : Tagging of the words, Data Collection of almost all the disease, We are regularly updating our vocabulary if the word is unfamiliar, WH- rules formation of factoid type questions, Dealt with compound sentences and simple sentences both, Incorporated the semantics, Mapping of unknown words to known words, Mapped with XML, Creation of SQL Query, Displaying the final results, All modules required are covered.

5. References

[1] Shen Song, Yu-N Cheah, Enya Kong Tang, Bali Ranaivo-Malancon*, Rule Extraction for Automatic Question Answering Based on Structural Clustering ,* International Journal of Computer Science and Network Security, Vol No.3, March 2008.

[2] Rashid Ahmad, Mohammad Abid Khan and Rahman Ali, *Efficient Transformation of a Natural Language Query to SQL for Urdu ,* Proceedings of the Conference on Language & Technology 2009.

[3] Mohammad Reza Kangavari, Samira Ghandchi,Manak Golpour, *A New Model for Question Answering Systems*, World Academy of Science Engineering and Technology 2008.

[4] Venkata Siva Rama Sastry K, Salil Badodekar, and Pushpak Bhattacharyya, *Questionto-Query Conversion in the Context of a Meaning-based, Multilingual Search Engine* ,Indian Institute of Technology Bombay, Mumbai.

[5] Vasin Punyakanok, Dan Roth and Wen-tau Yih, *Natural Language Inference via Dependency TreeMapping: An Application to Question Answering,* Association for Computational Linguistics, Volume 6, Number 9.

[6] Alessandra Giordani, *Mapping Natural Language into SQL in a NLIDB***,** University of Trento E. Kapetanios, V. Sugumaran, M. Spiliopoulou (Eds.): NLDB 2008, LNCS 5039, pp. 367–371, 2008. © Springer-Verlag Berlin Heidelberg 2008.