Intelligent Question Answering System using Natural Language Processing

Sandeep Pabolu, Pallepati Rakesh Reddy, Surparaju Lava Kumar, Mudunuri S S Sarat Chandra Varma

Abstract

As man is behind the quest for knowledge. This quest led us to the invention of lot many things including the most precious computers. This thrust starts with some questions like" Why?"," What"," Who" and so on. This finally led the working on question answering system. For such a system to work efficiently we need to consider the knowledge graph model for making the operation effective. The main objective of the project is to set up a code which generates approximate answers for user's quest based on the dataset. This sharing of ideas, experience and information should be available in the right place at the right time thereby reducing the user's wait time.

1. Introduction

For the best operation of the search engine in gathering the information from variety of sources is what server by the knowledge graph. In general, the words used are not simple the words they are the connection to real world entities which have some plot and some relation with a lot of other entities. All these entities are far related among them which if tracked down leads us to the web or graph which generates the knowledge graph. Type of information retrieval. Given a collection of documents, the system should be able to retrieve answers to questions posed in natural language.

The scope to ask a question in one's own way is what moves us forward. The main requirement is not to get related pages but the exact answer. This system has its cover everywhere science, education, assistants, etc. This makes it the most important issue where ever a computer has an application. So, working on the question answering system is worth exploring. The project aims in developing a code which could generate the answers for the questions given by the user. The answer data is searched through a specific data set. There by reducing the user's wait time and simultaneously decreasing the wait time.

Its work started in 1950 then the expanding took place in the mids of 1954 where the Russian sentences were converted to the English. But it was declared as failed project in 1966 which made the long suffering a sad ending leading to a complete tur over on the development site.

2.1 Knowledge Vault

Knowledge Vault will use the algorithm at mass-scale to form a single base of facts with the combination of all information including structured, unstructured and semi structured data. From the web. It is as effective as knowledge graph with organized information from the web.

Knowledge vault is larger than the knowledge base. It uses multiple extraction from different types of sources and various systems. Components used in the Knowledge graph are as follows: Extractor, Graph-base prior and knowledge fusion. It has various types of extraction methods such as txt, HTML tree(DOM), HTML Tables, Human Annotated Pages. For the link predictions, they have used the Path Ranking Algorithm(PRA), Neural Network Model(NNM).

These are used in a way whether they have a set of existing edges and will predict which other edges are having chance to exist. PRA is used for set-of-pairs of various entities that are connected to each

other. This will perform the random walks from beginning of source nodes. It is formed in the form of pairs. Neural network will view the link prediction problem compares to the matrix form.

Local Closed World Assumption(LCWA) will be used to approximate the truth. One of the sources foe Knowledge vault will be the Freebase where most of the information is extracted from it whereas the other data sources will be YAGO and DBpedia. It uses the fixed ontology-schemes such as Deep dive, NELL, Prospera and Read the web. Probase is a constructor used here for hierarchies. The basic difference between Knowledge vault and present approach is here, we combine facts extracted from data with the previous knowledge graph. It can have the uncertainty from the facts it has derived.

Knowledge vault can be improvised by the mutual exclusion from different tweets and the correlation between the facts [1]. The values can be represented from the multi-level abstraction, Dealing from the correlated data sources. There will be temporary facts which are true for only certain limits. It can be improved by adding various entities and relations in the graph.

2.2 Clustering

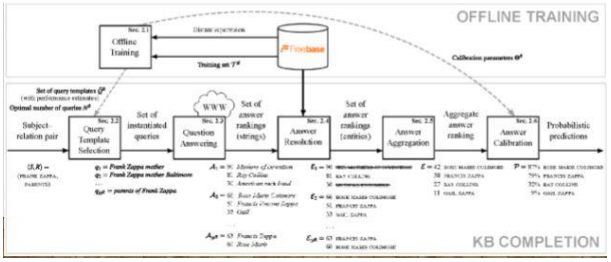
It is a platform for the knowledge gainers to push questions and get the answers in the form of comments, voting, and ratings. It gives permissions for a real time online communication where the questions and their discussions from the Q/A pairs. Its major application has been done by yahoo in the real-world applications. The question similarity has been tackled by three ways namely: lexical, syntactic, and semantic.

The methods used or the processes applied are Bow method, syntactic method, fuzzy matching based approach [2]. The algorithm goes as choosing the doc length, then followed by the occurrence of the subject in the data and finally setting up the probability of the word. The topic clustering approach is used for direct estimation of the weight of the word in the documentation:

$$w(t_{ij}) = \frac{h_k(t_{ij})}{\max_{t_{ij} \in q_i} h_k(t_{ij})}$$

The current system can be improved by using the topic based semantic similarity computing and unsupervised machine learning will improve the existing CQA systems. Ground truth represents the accuracy of the training sets. It twenty questions from the training sets. The experiment goes on as bow and cluster as bc, bow and cluster and filter bcf, updated bow and cluster upbc.

Syntactic tree matching for the state of art approach. Lda and cluster with our proposed questions. lda and cluster and filter for integrating purpose. Hence the paper is easy and short. It briefly explained the lds approach for topic modelling. The proposed model covered all the similarity computing measures lexical, t-distribution, t-weight.

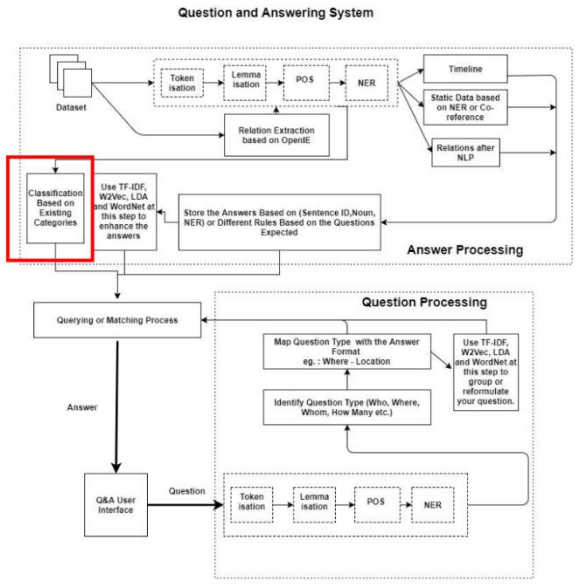


The process goes as named entity recognition followed by the replacement of the q with a place holder, running question answering system to get the answer entity. finally increasing the count of R,q. when too many queries are asked then there arises the computational challenging. Here the cpu time and the data base lookups are observed.

Heat maps are used to encode the average quality of queries starting from the template using the color code. Quality is measured in terms of mean reciprocal ranking. For dealing with the answers in the freebase various entities are linked to the answers. By this the type of answer we are searching is obtained very quickly. Through the

answer aggregation the output scores are converted into the percentages. Queries have many answers thus have probability range from 0-1. Because of using this aggregation, the best choice will be greedy, multiple queries is harmful, open vs. closed relations.

Architectural Diagram:

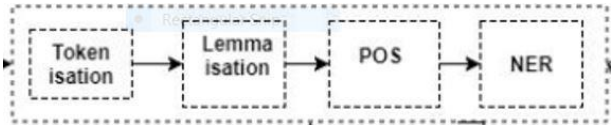


Tf-IDF:

Term frequency and inverse document frequency is a by default counter indicator program. It identifies the number of times a word is encountered in the selected document and the at the same time has the ability to list out the documents in which the particular word in used along with the name of the document.

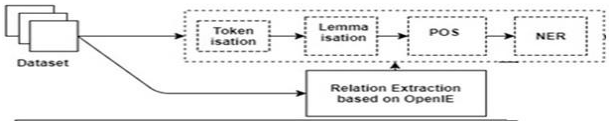
Natural language Processing (NLP):

NLP is a field in computer science engineering where the given set of words or the group of statements which are converted from human language to machine perception form. With the help of NLP we have applied lemmatization, parts of speech tagging, parsing, sentence breaking, stemming, word segmentation, terminology extraction which make the conversion from the normal data set to knowledge graph more effective and efficient.



Open IE:

Open Information Extraction is basically machine-readable presentation of the information text in the form of triplets. This feature makes the searching for an entity more reliable and more exact in figuring out the answer from the large data set.



Ontology:

It basically is a set of data in any format having the related data and information at the same place. This relation is made in such a form that it resembles a spider web like structure. Each and every entity has a relation in the ontology based in the data set.

DL Query:

These are used for testing the relationships or object properties between different classes. They can create class membership without the placeholders.

OWL

It is basically a language which was developed for presenting the tough knowledge of things, their relations, and expressions between different things.

Significance

The major significances can be listed out as under:

1. It increases the interactivity of the user with the database.
2. Searched answers can be viewed often.
3. Sentiment analysis can be performed for the better understanding.

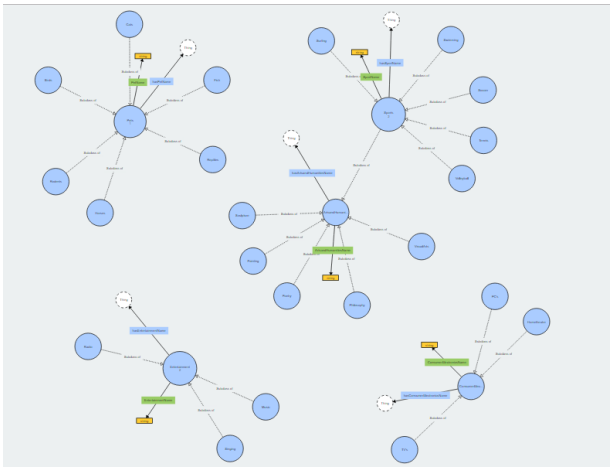
A proper usage of the linguistic resources.

Domain, Datasets, Ontology:

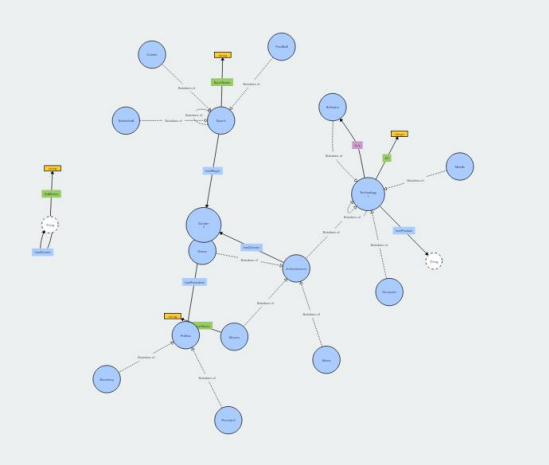
The data sets that we used for this project are available in the links followed. Basically, we have used two datasets in this project one of which is the news dataset and the other being sports. The links to the datasets are posted here.

SQuAD: Stanford Question Answering Dataset, it has Questions from the crowd workers on a set of Wikipedia articles. Answer of each question is from the corresponding passage. This dataset is much larger than the other datasets.

Yahoo Answers dataset: This dataset has 189467 QA pairs from top 20 categories from yahoo answers website. It also has total of 280 sub categories.



We have used different datasets from yahoo and BBC. Which holds datasets of different categories namely, sports, entertainment, politics, technology, Art-sand Humanities etc. We have taken sports, politics and entertainment datasets and constructed questions based on the dataset and processed the questions and dataset using NLP techniques, OpenIE, WordNet, Machine Learning to enrich the questions and datasets. Then the enriched datasets can be used to perform Q/A tasks. The main importance of enrichment is that system can be improved. the accuracy, quality and performance of the Q/A



Evaluation method:

1. We have used Natural Language processing (NLP) for tokenization, lemmatization, POS tagging, NER and Co-referencing.
2. Used Stanford Core NLP for processing the selected dataset.
3. The dataset used is related to Yahoo sports data on a match between Australia and New Zealand.
4. Processed the input data taken from a text document and stored the information in different text files.
5. Once after processing the question, we have used the data stored the output text files for answering them.
6. We have used the techniques TF-IDF to decide the prominent words in the dataset along with their scores and used Word2Vec to determine the similar words for the prominent words which will better help while understanding the questions asked and about how to answer them efficiently.
7. We have used the techniques OpenIE, WordNet, Clustering and Classification techniques to enrich the input dataset and Question Answering System.

Technologies Used: Java, Spark and Scala.

Triples Generation:

```
[(Greene,sets,sights,1.0), (Greene,sets sights on,world title,1.0)]
[]
[(He,had,settle,1.0), (He,had,settle in Greece,1.0), (He,settle behin
 American Justin Gatlin,1.0), (He,had,settle behind fellow Justin Gat
 American Justin Gatlin,1.0), (He,had,settle in Greece behind fellow
 (He,had,settle for bronze,1.0), (He,had,settle for bronze behind fel
 for bronze behind fellow American Justin Gatlin,1.0), (He,had,settle
 Gatlin,1.0), (He,settle behind,fellow Justin Gatlin,1.0), (He,had,se
 bronze in Greece behind American Justin Gatlin,1.0), (He,had,settle
 Greece,1.0)][][(It,was,my mistake,1.0)][(Greene,races in,Birmingham,
 0.02 seconds,1.0), (Greene,crossed,line,1.0)][(his title,is in,semi-
[(I,was running,all,1.0), (I,was running alone,all,1.0)][(I,was in,mi
 Lewis-Francis again at Friday 's Norwich Union Grand Prix,0.40698801
 again,0.40698801184380096), (Kansas star,is set,go with Lewis-Franci
 Grand Prix,0.40698801184380096), (Kansas star,is set,go with Lewis-F
```

SPARQL Queries:

Active Ontology × Entities × Individuals by class × SPARQL

SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?subject ?object
        WHERE { ?subject rdfs:subClassOf ?object }
```

SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX onturi: <http://www.semanticweb.org/shali/ontologies/2017/6/(
SELECT ?University
        WHERE {?University onturi:playsGame onturi:AmericanFoo
```

SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX onturi: <http://www.semanticweb.org/shali/ontologies/2017/6/Obama#
SELECT ?subject
        WHERE { ?subject onturi:isPresidentOf onturi:UnitedStates}
```

Concerns and Issues:

- Notable issues are about processing data for some set of questions.
- Filtering stop words, synonyms. Co-referencing for complex datasets.
- Relation between the persons in the text.
- Related words questions need to be achieved.

Future Work:

Furthermore, we can extend our current questions set to much more complex set. Also, Processing data using Ontology with much more correct way. We need to implement the relation analysis, enrich the question type with yes/no answers in the knowledge graph. Full-Fledged Implementation of the Knowledge Graph for the Dataset.

Results:

Sample K-means-Yahoo Dataset:

```
Corpus summary:
     Training set size: 10 documents
     Vocabulary size: 82259 terms
     Preprocessing time: 77.803108814 sec

Finished training KMeans model.  Summary:
     Training time: 0.67535386 sec
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/dataset/BusinessAndFinance.Other-Taxes.txt;0
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/dataset/ComputersAndInternet.txt;0
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/dataset/EntertainmentAndMusic_magazines.txt;0
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/dataset/EntertainmentAndMusic_radio.txt;0
```

K-Means-BBC-Sport:

```
Corpus summary:
     Training set size: 25 documents
     Vocabulary size: 5917 terms
     Preprocessing time: 333.225723279 sec

Finished training KMeans model.  Summary:
     Training time: 1.212014935 sec
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/athletics/001.tx
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/athletics/002.tx
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/athletics/003.tx
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/athletics/004.tx
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/athletics/005.tx
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/cricket/001.txt;
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/cricket/002.txt;
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/cricket/003.txt;
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/cricket/004.txt;
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/cricket/005.txt;
file:/C:/Users/USER/Documents/Spark_MachineLearning/data/bbcsport1/football/001.txt
```

Final Results after LDA:

```
Corpus summary:
     Training set size: 417 documents
     Vocabulary size: 854 terms
     Training set size: 2423 tokens
     Preprocessing time: 3.801622757 sec

Finished training LDA model.  Summary:
 Training time: 6.934879554 sec
 Training data average log likelihood: -95.

5 topics:
TOPIC_0;connection;0.030922248292319265
TOPIC_0;pragmatic;0.02690878152282403
TOPIC_0;Originator;0.02443295327258558
TOPIC_0;check;0.022290036662386385
TOPIC_0;input;0.020578378303486064
TOPIC_0;character;0.019718375317755072
TOPIC_0;wide;0.017389396600966833
TOPIC_0;load;0.016898979702795396
TOPIC_0;Pretty;0.014923624938546124
TOPIC_0;soon;0.014731449663492822
TOPIC_0;University;0.014220648196245527
TOPIC_0;convince;0.014039842892520141
TOPIC_0;interpreter;0.014014099385562674
TOPIC_0;therefore;0.013990681689260381
TOPIC_0;computation;0.013563021169328254
```

Accuracy:

Naïve-Bayes:

```
Confusion matrix:
1.0  0.0  1.0  0.0  1.0
0.0  1.0  0.0  0.0  2.0
0.0  0.0  1.0  0.0  3.0
0.0  0.0  0.0  3.0  3.0
0.0  1.0  0.0  0.0  4.0
Accuracy: 0.533333333
```

Set of questions and answers:

What sport is discussed in the dataset? *Cricket.*

How many players are playing from Australia? *12*.

Who has won the match? *Australia*.

When was the match played? *May 10, 2015*

Which player has scored most number of runs? *Hayden*.

Are there any players injured in the match? *No*.

Is there any possibility of Australia winning the match based on the commentary? *Better Chances for winning the match*.

References

- Paulheim, Heiko. "Knowledge graph refinement: A survey of approaches and evaluation methods." Semantic web 8.3 (2017): 489-508.
- Zhang, Wei-Nan, et al. "A topic clustering approach to finding similar questions from large question and answer archives." PloS one 9.3 (2014): e71511.
- Dong, Xin, et al. "Knowledge vault: A web-scale approach to probabilistic knowledge fusion." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.
- Zhang, Ce, et al. "DeepDive: declarative knowledge base construction." Communications of the ACM 60.5 (2017): 93-102.
- Collarana, Diego, et al. "Semantic data integration for knowledge graph construction at query time." Semantic Computing (ICSC), 2017 IEEE 11th International Conference on. IEEE, 2017.
- Wu, Sen, et al. "Fonduer: Knowledge Base Construction from Richly Formatted Data." arXiv preprint arXiv:1703.05028 (2017).
- Duan, Weiwei, and Yao-Yi Chiang. "Building knowledge graph from public data for predictive analysis: a case study on predicting technology future in space and time." Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data. ACM, 2016.
- Speer, Robert, Joshua Chin, and Catherine Havasi. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." arXiv preprint arXiv:1612.03975 (2016).
- Lehmann, Jens, et al. "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia." Semantic Web 6.2 (2015): 167-195.

- Andrieu, Olivier (14 January 2014). "Le Knowledge Graph de Google ferait baisser le trafic de Wikipedia" (in French). Abondance. Retrieved 15 January 2014.
- Vesselin Petrov (2011). "Chapter VI: Process ontology in the context of applied philosophy". In Vesselin Petrov, ed. Ontological Landscapes: Recent Thought on Conceptual Interfaces Between Science and Philosophy. Ontos Verlag. pp. 137 ff. ISBN 3-86838-107-4.
- Barry Smith: Objects and Their Environments: From Aristotle to Ecological OntologyThe Life and Motion of SocioEconomic Units (GISDATA 8), London: Taylor and Francis, 2001, 79-97.
- Banko, Michele; Cafarella, Michael; Soderland, Stephen; Broadhead, Matt; Etzioni, Oren (2007). "Open Information Extraction from the Web". Conference on Artificial Intelligence.