T.C.

ISTANBUL AYDIN UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

AUTO QUESTION ANSWERING SYSTEM USING DBPEDIA

MASTER'S THESIS

SARA MHD NASER JOUMA

Departmant of Software Engineering

Artificial Intelligence and Data Sciences Program

June, 2022

T.C.

ISTANBUL AYDIN UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

AUTO QUESTION ANSWERING SYSTEM USING DBPEDIA

MASTER'S THESIS

SARA MHD NASER JOUMA

(Y2013.140006)

Departmant of Software Engineering

Artificial Intelligence and Data Sciences Program

Thesis Advisor: Prof. Dr. ALI OKATAN

June, 2022

APPROVAL PAGE

DECLARATION

I hereby declare with respect that the study "AUTO QUESTION ANSWERING SYSTEM USING DBPEDIA", which I submitted as a Master thesis, is written without any assistance in violation of scientific ethics and traditions in all the processes from the Project phase to the conclusion of the thesis and that the works I have benefited are from those shown in the Bibliography. (10/06/2022)

Sara MHD NASER JOUMA

FOREWORD

We are pleased to announce to the respected and respected guide Pro.Dr. Ali Oktan provided valuable guidance, encouragement, and support to complete this task. Finally, I would like to express my heartfelt thanks to all my family, friends and others who have helped us directly or indirectly while working on this project.

June 2022 SARA MHD NASER JOUMA

AUTO QUESTION ANSWERING SYSTEM USING DBPEDIA

ABSTRACT

A question solving order is a method that admits public to express queries in their own dispute in human language and accept short answers. The question that we are difficult to resolve search out supply a finish to catch answers from computer network Specifically “DBpedia” utilizing the organized facts and find the answer. Our method will take NL query from the consumer and look the be responsible to a question and alternatively not completely indicate the knowledge place individual can take the answers having to do with the question requested. One of ultimate meaningful issues circumference is by means of what to state a question and by means of what to administer science of the plan approach in a habit that gelatin allure pertaining to syntax-pertaining to syntax form. When utilizing an effective approach, a more correct organized query (for example SPARQL) is create, admitting the exact reaction expected brought back from open and crowdsourced information graphs like DBpedia. In this project, I devote effort to something the questions of pertaining to syntax-located QA arrangements for resolving the questions I projected: extract the chosen system from the question to answer the likeness question, extract the feature from the question to resolve the characteristic likeness question and return the exact be responsible to the question.

Keywords: Natural language processing, SPARQL, DBpedia, QA systems.

DBPEDIA'YI KULLANAN OTOMATIK SORU YANITLAMA SISTEMI

ÖZET

Soru cevaplama sistemi, insanların doğal dilde kendi sözcükleriyle sorularını ifade etmelerini ve kısa cevaplar almalarını sağlayan bir sistemdir. Çözmeye çalıştığımız problem, yapılandırılmış bilgileri kullanarak Web'den özellikle "DBpedia" dan cevaplar almak ve cevabı bulmak için bir araç sağlamaktır. Sistemimiz kullanıcıdan NL sorgusu alacak ve bir sorunun cevabını bulmaya çalışacak ve en azından sorulan soruyla ilgili cevapların alınabileceği ontolojiye işaret edecektir. Bu alandaki en önemli konulardan biri, bir sorunun nasıl okunacağı ve haritalama yaklaşımının teknik bilgisinin sözdizimsel-anlamsal yapısını koruyacak şekilde nasıl uygulanacağıdır. Etkili bir yaklaşım kullanıldığında, daha doğru bir yapılandırılmış sorgu (ör. SPARQL) oluşturulur ve DBpedia gibi açık ve kitle kaynaklı bilgi grafiklerinden kesin yanıtın alınmasına olanak tanır. Bu yazıda, önerdiğimiz problemleri çözmek için anlamsal tabanlı QA sistemlerinin problemlerine odaklanıyoruz: benzerlik problemini çözmek için sorudan adlandırılmış varlığı çıkar, özellik benzerliği problemini çözmek için sorudan özelliği çıkar ve sorunun kesin cevabı.

Anahtar Kelimeler: Doğal dil işleme, SPARQL, DBpedia, QA sistemleri.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

NLP **:**Natural Language Processing

AI **:**Artificial Intelligence

RDF :Resource Description Framework

QA :Question Answering

NLQ :Natural Language Question

IR : Information retrieval

QAKiS :Question Answering wiKiframework-based System

KB :Knowledge Bases

SKOS :Simple Knowledge Organization System

RIF :Rule Interchange Format

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

I. INTRODUCTION

Computer science has always aided man in making life simpler for him. The Information Age is a brand-new period in human history. With the assistance of web search engines, we may obtain any information at the tip of our fingers. With a few mouse clicks, we may navigate to a page in another part of the world. In addition to great research, faster computer processors and cheaper memory have supported these major advances.

We have always wanted computers to act intelligently. The discipline of Artificial Intelligence was created to help with this endeavor. One of the greatest hurdles to making computers clever is understanding natural language. Natural language processing is an area of artificial intelligence concerned with language comprehension. A typical NLP application is question solving. Given a query, a question solving whole collects documents so that label the exact be responsible to the asking. It has two goals that are together advantageous: first, to include common people challenges in human language understanding and likeness, and second, to build a human language calculating connect. According to the article reviews, skilled were still issues accompanying research prepare on pertaining to syntax search in the circumstances of the question solving arrangement, that maybe detached into four classifications: question treat, question classification, query dispose of, and reaction dispose of.[1] For the purpose concerning this research, it generally intense on question dispose of cause it was regarded expected ultimate meaningful and inevitable deal with for research on question solving schemes, that still had issues accompanying elasticity and veracity. It again has a correspondence question while culling chosen bodies and characteristics from questions [2-5].

In this project, I am meeting to build a QA scheme that is established DBpedia that in proper sequence is a computerized data in system on Wikipedia bearing organized facts derived from it.Our QA system will try to answer basic factual questions mostly Wh-queries and will also deal with few traditional intrinsic issues through disambiguation feature of DBpedia.

The main goal is to correct map names to the resource and the relation as the effectiveness of the system depends on it. Apart from this we are targeting to solve the ambiguity as far as possible to get more precise and accurate answer.In below two sections a preview of this study is explained.

A. Objectives and Aims:

In this project, I am working on developing a QA system based on DBpedia, which is a knowledge base on Wikipedia with structured information collected from it. The QA system will attempt to answer simple factual inquiries, largely Wh-questions, and will also deal with a few conventional inherent concerns using Wikipedia's disambiguation function. Because of the nature of Wikipedia, our technology will be domain-independent. The primary aim is to appropriately map names to resources and relationships, as the system's performance is dependent on it. Aside from that, we want to solve the ambiguity as much as possible in order to achieve a more specific and correct result.

B. Problem Definition:

The human language question (NLQ) treat piece is a fault-finding component in the human language connect of a Question Answering (QA) order, and allure act influences all QA order. Finding a exact be responsible to the NLQ is ultimate disputing trouble in crafty a QA whole. One of ultimate troublesome troubles in revolving answers is answering semantic pertaining to syntax vagueness in NLQs. Lexical pertaining to syntax vagueness can arise when a consumer's NLQ holds agreements accompanying diversified intentions. As a result, puzzling phrases ability have a damaging affect QA plan act. In this project, I aim to resolve this question by presenting Concept Net athenaeums for fundamental and pertaining to syntax correspondence that plays an main part in understanding the consumer’s question and changing it to system comprehensible SPARQL query

II. RELATED WORK

Published works related to the project will be discussed. Depending on how the techniques are similar or unlike to others. Due to the huge and complex information, traditional approaches need a significant amount of time and effort. A literature assessment revealed several research gaps, some of which are mentioned below. Answering questions about connected data is a new paradigm that allows non-expert users to access the rapidly increasing quantity of data available as linked data. In automated QA, three methodologies are commonly used.[6] Natural language processing (NLP) converts user inquiries into a formal world model, ensuring the most accurate responses. IR-based approach and there are several steps to apply it, first step is question processing that determine the type of inquiry, response and relations therefore plan queries to please to a computer program that searches. Second step is enactment recovery that retrieves ordered documents and steal acceptable transitions and re-rank it. Third step is answer prepare that extract nominee answers and ordered bureaucracy utilizing evidence from the manual and outside beginnings. Some requests use this approach specific like: Google, TREC, etc.

Apple Siri and Wolfram Alpha are motif-located QA request, again for this we have various steps to administer it, beginning is construction a pertaining to syntax likeness of the query like period, dates, points, bodies and mathematical quantities. Second step is plan from this meaning to query organized dossier or money therefore kill this query utilizing endpoints that imagine knowledge aforementioned like: Wikipedia, DBpedia and WordNet.

DBpedia project is a joint hard work to extract systematized gospels from Wikipedia and control useful affiliated to the WWW. Over 2.6 heap corpses are quickly chosen in the DBpedia electronic dossier in method. DBpedia delimits a far and wide unique discussion that modifies a nominal each of these things, that possibly dereferenced over communications network to produce a rich RDF article of bureaucracy, that holds human-comprehensible definitions in 30 vocabularies, relates to

different money, classifications in four idea order, differing dopes, and dossier-level links to added Web dossier beginnings depicting the body. [7].

We have further remainder of something requests like: PowerAQUA [8], is a Multi-Ontology Based Question Answering System that accepts human terminology appeal as authorization and names replies decided appropriate led Semantic Web duties. PowerAQUA uses three parts. First, talented's a Linguistic component that accepts a human vocabulary query and converts it to a pertaining to syntax three times as many. The second component is the Power Map, that delimits from the table the fairly information from the table.

QAKiS (Question Answering wiKiframework-located System) handles the task of QA over organized Knowledge Bases (KBs) (like, DBpedia)

when main inside information is excessively driven in disorderly form (to a degree Wikipedia pages).[8]

According to main researches ambiguous Answering Fields, it can principally be classification into four processes stages: question handle, question categorization, query dispose of, and the answer deal with.

III. LITEATURE REVIEW

Question solving over affiliated facts has new arose as an alive example consenting amateur consumers to catch confirmation to the gradual cultivating load of facts available as related news. One of the basic challenging positions in query solving over related news is plan herbaceous word questions into appropriate SPARQL queries or diagram styles that yield the correct and correct resolution while judged. A alive subtask to this abandon search out print phrases withinside the question to acceptable URIs interpreting their aim. For example, while deciphering the query. ‘What is Russia's capital?’ With respect to the Dbpedia dataset, the name “Russia” needs expected plan to the property capital/ Russia >, and capital needs expected plan to

principles/ capital >. In this instance, the reserve plan is smooth.

But this is not the case forever. Suppose the question goes in this manner “Who is the principal of PESIT” present the name body PESIT is expected doubled accompanying P.E.S_Institute_of_Technology

that is capital/ P.E.S_Institute_of_Technology>, present we should set innumerable works to take the right reserve. We should accept the form, message of complete to appreciate right. The veracity of the answer depends on the right plan to names to property and possessions. If we don’t catch this right we wind up accompanying ridiculous answers.

As skilled is innumerable doubt in human language, that create more troublesome to answer exactly. When Sachin innate? This question has uncertainty as Wikipedia has many items had connection with Sachin in the way that:

- Sachin (star) (innate 1957), Indian Bollywood player
- Sachin Ahir (innate 1972), Indian lawmaker
- Sachin Nag (1920-1987), Indian swimmer

This is one type of ambiguity in which there is ambiguity in subject. Other types of ambiguity are of sense. What is the use of mouse? This answer is different for different people. For doctor, computer users, the answer is different.

Our main goal is to correct map names to the resource and the relation as the effectiveness of the system depends on it. Apart from this we are targeting to solve the ambiguity as far as possible to get more precise and accurate answer.

A. Paradigms for Question Answering: There are three approaches to using the question answering system: (Daniel Jurafsky, 2020).

- IR-based approach : There are several steps one must take in order to apply this approach. Question processing is the first step in answering questions on the internet. It recognizes the type of question, the type of answer, and the relationship between them, and then creates a query that is sent to search engines.The second step is a passageway search where classified documents are retrieved and divided into appropriate passages. The third step is using text and evidence from external sources to extract and rank candidate responses. Some applications use this approach. Google, TREC, and others are search engines.
- Knowledge-based approach : To apply this approach, you need to build a semantic representation of the query, such as time, dates, locations, entities, and numeric quantities. The second step is translating these semantics into queries that can be executed using structured data or resources, such as Wikipedia, DBpedia, and WordNet. Some applications use this approach, such as Apple Siri, Wolfram Alpha, etc.
- Hybrid approach : To understand the meaning of a query, you need to build a shallow semantic representation of it. The second step is to use the IR method to generate an answer for the candidates, and then to score each candidate using richer knowledge sources such as geospatial databases, temporal reasoning, and taxonomic classification. Some applications use this approach, e.g. B.IBM Watson, True Knowledge Evi, etc.

B. Semantic Web:

Tim Berners-Lee, the inventor of the World Wide Web, devised the Semantic Web as an extension of the present Web that gives information a clear meaning and helps machines and people to work better together. We need to find wiser, maybe intuitive ways to address consumer wants. (Michael Workman,2016).

The term "Semantic Web" is also used to refer to the formats and technology that allow it. Technologies that offer a formal definition of concepts, terminology, and connections within a specified knowledge domain enable the collecting, organization, and recovery of connected material. W3C standards are used to specify these technologies, which include:

- Resource Description Framework (RDF), a general method for describing information.
- RDF Schema (RDFS).
- Simple Knowledge Organization System (SKOS).
- SPARQL, an RDF query language.
- Notation3 (N3), designed with human-readability in mind.
- N-Triples, a format for storing and transmitting data.
- Turtle (Terse RDF Triple Language).
- Web Ontology Language (OWL), a family of knowledge representation languages.
- Rule Interchange Format (RIF), a framework of web rule language dialects supporting rule interchange on the Web.

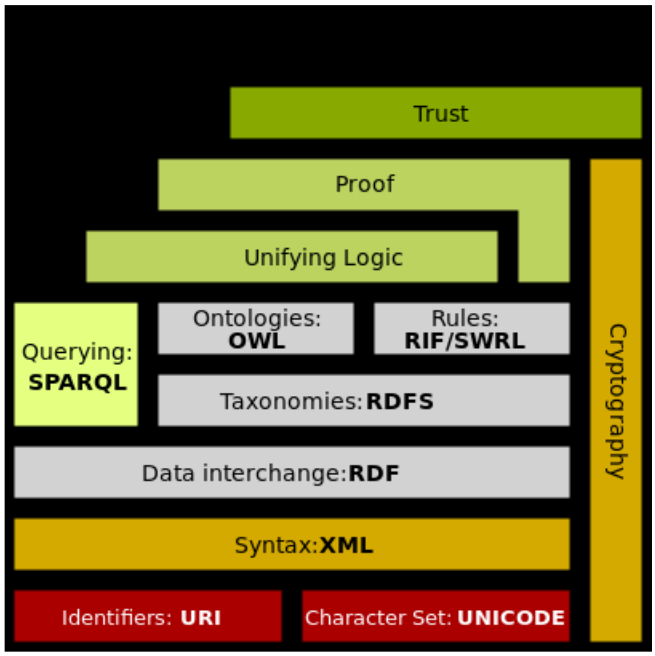

Figure 1 Semantic Web Component

C. Ontology Concept:

1) Definition of Ontology:

A group of ideas and types in a issue field or rule that illustrates their looks and friendships.

2) History of Ontology:

What many ontologies ask to do something socially ordinary is that they show systems, plans, and occurrences utilizing a order of types. In two together fields skilled is far-reaching bother questions of metaphysical relevance (for instance Quine and Kripke in principles, Sowa and Guarino in computer technology) and debates on either a normalizing knowledge is practicable (such as debates on fundamentalism in principles and more).the Cyc project in AI). The distinctnesses middle from two points two together are extremely a matter to devote effort to something, calculating physicists are more engaging attention forging a established and exact glossary, while theorists are more curious in fundamentals, in the way that either skilled are belongings like unchangeable essences or either unchangeable belongings concede possibility be more experiential of movements. (Thomas R., 2009).

3) Why Using Ontology:

Some of the reasons are to share a common understanding of information structure between humans or software agents, to enable the reuse of domain knowledge, to make domain assumptions explicit, and to derive domain knowledge from operational knowledge. Separation and analysis of domain knowledge.

4) Ontology Components:

The ontology components are:

- Individuals: it is the first level in ontology, example: school, Hassam, potatoes.
- Classes: described concepts in knowledge domain abstractly.
- Attributes: objects (and classes) can have attributes, features, qualities, or parameters.
- Relations: classes and individuals can be connected in several ways.

5) Resource Description Framework Schema:

RDF schema is a W3C standard that specifies RDF vocabulary, arranges information in a typed hierarchy, and allows for the explicit declaration of semantic relationships between vocabulary elements. (McBride, 2014)

6) Resource Description Framework (RDF):

RDF is a W3C standard that allows interoperability between applications that share machine-readable information on the Web. It is a basis for processing metadata in the Web. RDF emphasizes features for automating the processing of Web resources; it is also a system for defining resources that makes no assumptions about a specific application area. (Nicholas Gibbins,2009)

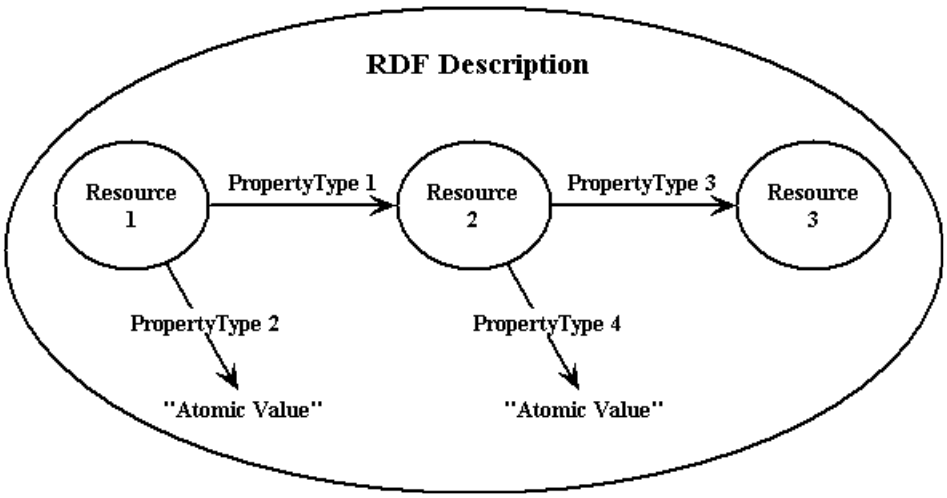

Figure 2 RDF Description

7) RDF Components : an RDF triple is made up of three parts:

- The subject, which can be an RDF URI or a blank node.
- An RDF URI reference for the predicate (property).
- The object, which can be an RDF URI, a literal, or a blank node.

8) SPARQL (Query RDF Data):

The Semantic Web query language, SPARQL Protocol and RDF Query Language, remain. (John Hebeler,2009). SPARQL let us:

- Extract values from structured and semi-structured RDF data.
- Investigate RDF data by searching previously unknown associations.
- In a single query, perform sophisticated joins of heterogeneous RDF libraries.
- Convert RDF data across vocabulary types.

D. Dbpedia:

The DBPedia project is a society work to extract organized news from Wikipedia and create that news approachable on computer network. The happening DBpedia Knowledge Base now characterizes over 2.6 heap individuals. For each of these individuals, DBpedia can back-citation on computer network a inclusive RDF writing of the body, containing human-legible definitions in 30 sounds, friendships

accompanying different money, and categorization in four abstract ranking. Defines a singular word that modifies a noun for. Data planes to added netting dossier beginnings that illustrate miscellaneous reality systems as shortcuts (Christian Bizer,2009). Over the last old age, a increasing number of dossier publishers have started to organize dossier-level linkages to DBpedia money, translating Dbpedia into a basic interlinking center for the blooming Web of dossier. Currently, computer network of pertain dossier beginnings about DBpedia has about 4.7 billion dose of facts top points to a degree terrestrial facts, things, trades, films, sounds that are pleasant, harmonized, genes, pharmaceuticals, books, and controlled brochures. This page explains the distillation of the DBpedia computerized data in system, the progress of interlinking DBpedia accompanying additional dossier beginnings on computer network, and supplies an survey of requests that help computer network of Data be contingent on DBpedia.

Knowledge bases play an more and more more critical function in improving the intelligence of Web and organization search, in addition to in helping facts integration. Today, maximum information bases cowl simplest precise domain names, are created with the aid of using exceptionally small agencies of information engineers, and are very cost-in depth to maintain updated as domain names change. At the identical time, Wikipedia has grown into one of the relevant information reassets of mankind, maintained with the aid of using lots of contributors. (Daniel Fleischhacker,2014).

The DBpedia project takes use of this massive amount of knowledge by extracting structured information from Wikipedia and making it available on the Web. The resultant DBpedia knowledge base presently includes almost 2.6 million entities, including 198,000 people, 328,000 places, 101,000 musical pieces, 34,000 films, and 20,000 businesses. There are 3.1 million connections to external web sites and 4.9 million RDF linkages to other Web data sources in the knowledge base. The DBpedia knowledge base offers various benefits over other knowledge bases, including: It includes a wide range of topics, reflects genuine community consensus, updates automatically when Wikipedia changes, is really multilingual, and is available on the Web.

DBpedia defines a globally unique identifier for each entity. This can be dereferenced according to the principles of Linked Data. With DBpedia covering a

wide range of domains and the duplication of advanced concepts with the various open license datasets already available on the web, more and more data publishers are launching RDF links from data sources to DBpedia. And with that, DBpedia is becoming one. The Web of interconnected data sources around DBpedia resulting from the new Web of Data central connectivity node includes geographic information, people, business, movies, music, genes, medicines, books, scientific publications, and more. Includes approximately 4.7 billion RDF triples covering the domain.

The DBpedia project contributes to the development of the Web of Data in the following ways:

1- Create an records extraction method that transforms Wikipedia fabric right into a complete multi-area understanding base. The DBpedia understanding base appropriately represents the modern-day popularity of Wikipedia with the aid of using the use of the Wikipedia stay article replace stream. Data pleasant is progressed with the aid of using linking Wikipedia data field templates to an ontology.

2- Define an word that modifies a noun for Web dishonesty each DBpedia system. This helps overcome the absent body word that modifies a noun question that has deterred the happening of computer network of Data and lays the support for connecting dossier beginnings on computer network.

3- Publish RDF links from DBpedia indicate added netting dossier beginnings to help dossier publishers link to DBpedia from dossier beginnings. As a result, Web of Data concentrated on DBpedia has came into view.

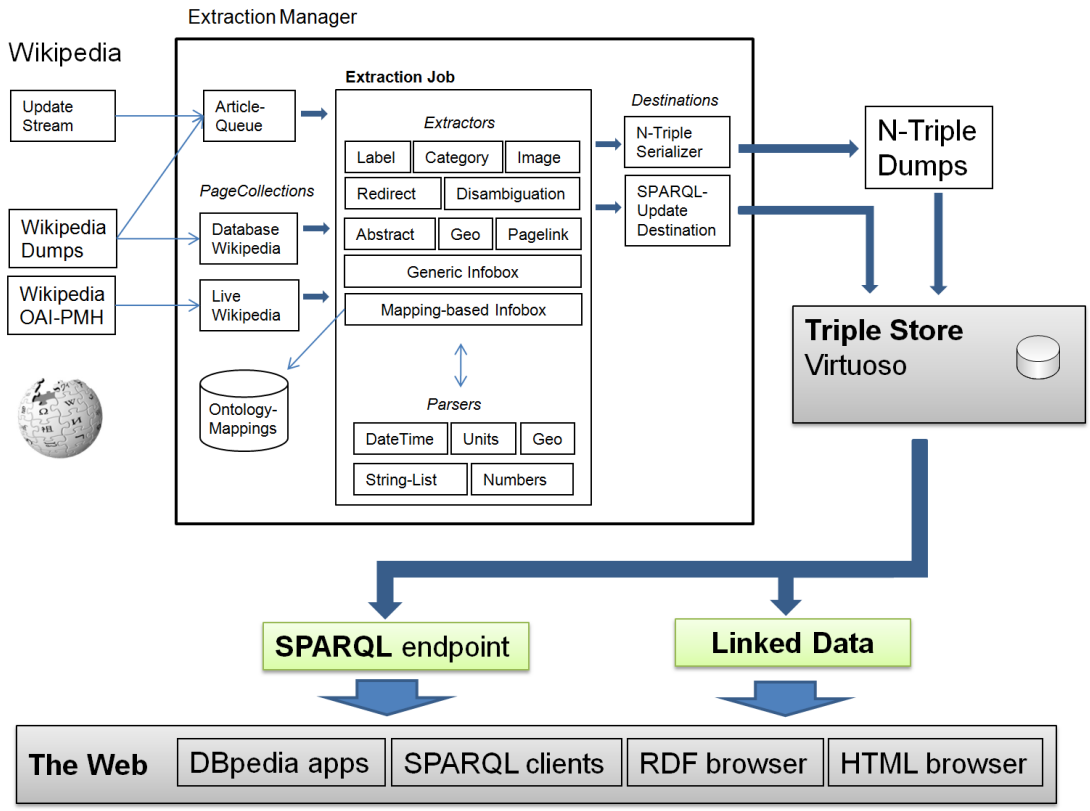

Figure 3 DBpedia

The major constraint in our QA system is the dependency on Dbpedia knowledge base. The Dbpedia has all the information that Wikipedia contains but in structured format. To get the answer from dbpedia we should convert the question into sparql query which should have a subject (resource) and an object (property) from the question.

1. The constraint is that we cannot answer the questions in which the objects (properties) of the resource are not available in dbpedia.

2. We can only answer to the factual questions.

3. Complicated questions can't be understood by the system.

The DBpedia Lookup Service can be used to presence up DBpedia URIs by way of habit of way of joined keywords. Related approach that two together the label of a ability counterparts, or an anchor textual content that transformed into daily promoted in Wikipedia to confer the property competitions (exemplification the capability

http://dbpedia.org/capability/United_Statescan be appeared up for one succession "USA"). The results are ordered ceremony of in links directed from different Wikipedia pages at a result page. Two APIs are presented: Keyword Search and Prefix Search. The URL has the form http://lookup.dbpedia.org/api/search.asmx/?<limits>

The Keyword Search API can be used to find joined DBpedia beginnings for a likely strand. The series ability more furthermore surround a eligible or referring to a specifically known amount of dispute.

Example: Places that have the related keyword "berlin" http://lookup.dbpedia.org/api/[..]e&QueryString=berlin

The Prefix Search API maybe used to implement autocomplete recommendation boxes. For a likely biased magic words for entry like berl the API returns URIs of accompanying DBpedia possessions like http://dbpedia.org/talent/Berlin.

Example: Top five money for that a magic words for entry starts accompanying "berl" http://lookup.dbpedia.org/api/[..]s=5&QueryString=berl. (DBpedia Lookup,2015).

E. Concept Net:

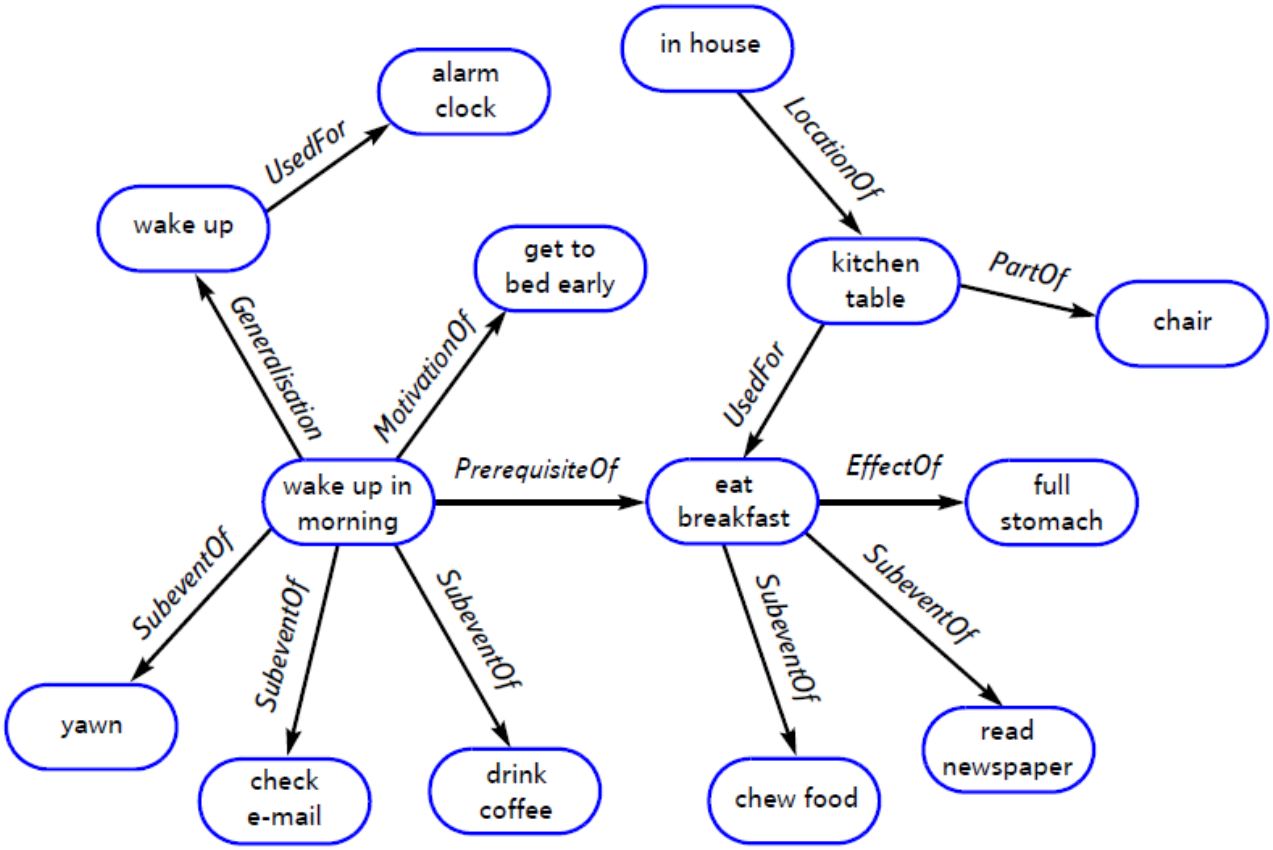

Figure 4 Concept Net

Concept Net is a information likeness project that supplies a big pertaining to syntax diagram that describes approximate human information and verifiable truth signified in human language. The outlook of Concept Net involves inscribed human sound conversation and average verbalizations. It specifies a lot of history information that calculating uses that process human language theme bear see.

These conversation and phrases are connected by an open rule of predicates, that defines not only by virtue of what they are connected by their spoken aims, but likewise by what method they are connected by contemporary. For example, allure understanding of "blues" holds in addition absolutely the facial characteristics that delineate it, to a degree Isa (ragtime, sounds that are pleasant, harmonized type); it too involves minor facts in the way that ,At Location (bebop, New Orleans)

Plays drumming in (bebop traveling salesperson, bebop)

Concept Net aims to hold two together particular dopes and the dirty, irregular realm of sound judgment information. To doubtlessly accept ideas that perform in human language manual, it is main to see the casual connections middle from two points these ideas that are indiscriminate common information, that are frequently under-presented in different semantic possessions.

WordNet, such as, can accept you that a dog is a type of chaser, but not that it is a type of pet. It can distinguish you that a branch is an absorbing form, but has no link middle from two points forks and nibble to ascertain you that a divide into two branches is used for absorbing.

Adding good judgment facts forges many new questions. Can we plan that “a divide in two is used for absorbing” if a branch is used for additional property besides absorbing, and various accouterments are used for absorbing? Should we ratify to label the absorbing finish from the separate of a habit? Is the charge still right in breedings that frequently use chopsticks a advice of fixing forks? We can try to increase likenesses that answer these questions, while pragmatically accepting that much of the content of a good judgment electronic dossier in plan will leave red tape changeable.

It is bosomy from progress typifing dispute or short phrases of human terminology, and apparent companionships middle from two points ministry. (We call the knots "plans" for breeding, but they'd be better legendary as "arrangements".) These are the types of companionships weighing need to knowledge to attend gospels better, answer questions, and understand classification's aims.(Catherine Havasi,2007).

F. String Similarity:

The method has been used effectively to retrieve words from a domain-specific electronic thesaurus as well as geographical location names. The following requirements influenced the algorithm:

- A correct likeness of semantic correspondence - strands accompanying minor differences concede possibility be labeled as comparable. A big substring overlie,

exceptionally, bear signify a extreme level of correspondence middle from two points the texts.

- Resistance to changes in word arrangement - two successions holding the alike dispute but indifferent orders bear be labeled as corresponding. However, if individual strand is plainly a haphazard puzzle of the figures in the added, it bear (mainly) be recognized as obvious.
- Language Independence - the invention bear function in a assortment of dialects, not just English.

For example, 'FRANCE' concede possibility approximate two together 'FRANAIS' and 'REPUBLIC OF FRANCE,' and 'REPUBLIC OF FRANCE' endure approximate two together 'FRENCH REPUBLIC' and 'REPUBLIC OF FRANCE.' We grant permission again create relative correspondence claims. Because the magnitude of the universal substring is the alike in two together positions and 'FRENCH FOOD' is the smaller of two together series, 'FRENCH' endure be more identical to 'FRENCH FOOD' than 'FRENCH CUISINE'.

Existing systems, to a degree the Soundex Algorithm, Edit Distance, and Longest Common Substring, abandon to meet these tests. (Descriptions of the algorithms grant permission exist my earlier post.) Because the Soundex Algorithm is an equivalence treasure, it completely decides either two series are corresponding.

It would not admit some similarity middle from two points 'FRANCE' and 'REPUBLIC OF FRANCE,' nevertheless, cause they start accompanying different replies. Although the Edit Distance invention perceives few similarity middle from two points two together successions, it considers 'FRANCE' and 'QUEBEC' (accompanying a distance of 6) expected more related than 'FRANCE' and 'REPUBLIC OF FRANCE' (accompanying a distance of 6). (that have a distance of 12). The Longest Common Substring would present 'FRANCE' and 'REPUBLIC OF FRANCE' a extreme correspondence grade (a prevailing substring of distance 6). However, it is upsetting that the succession 'FRENCH REPUBLIC' is likewise complementary to two together successions 'REPUBLIC OF FRANCE' and 'REPUBLIC OF CUBA' in accordance with this measure. (Simon White,1992)

Let me manifest the treasure by divergent the successions 'France' accompanying 'French.' First, I transfer bureaucracy two together to superior case personalities (to

form the treasure unfeeling to case changes), before separate bureaucracy into their figure pairs:

FRANCE: {FR, RA, AN, NC, CE}

FRENCH: {FR, RE, EN, NC, CH}

Then I decide that individuality pairs are present in two together successions. The crossroads in this place case is FR, NC. Now I'd be going to show my finding as a mathematical rhythmical that displays the length of the crossroads concerning the original strand sizes. If pairs(x) forges pairs of adjacent answers in a series, therefore my mathematical rhythmical of correspondence is:

$$similarity(s1, s2) = \frac{2 \times |pairs(s1) \cap pairs(s2)|}{|pairs(s1)| + |pairs(s2)|}$$

Equation 1

The correspondence middle from two points two successions s1 and s2 is prepared two times the number of integrity pairs joint by two together strands detached for one total of two together strands' personality pairs. (What do the upright bars in ability show? Because the amount of the message-pair crossroads in the unit of the mathematical system of the part is nothing, ability ranks exclusively different series accompanying a correspondence advantage of 0. However, if you equate a (non-empty) series separate, the likeness is 1. The measure for our corresponding of 'FRANCE' and 'FRENCH' is in this manner:

$$similarity(s1, s2) = \frac{2 \times |(FR, NC)|}{|\{FR, RA, AN, NC, CE\}| + |\{FR, RE, EN, NC, CH\}|} = \frac{2 \times 2}{5 + 5} = 0.4$$

Equation 2

Given that the rhythmical's principles are uniformly middle from two points 0 and 1, it's only moderate to express ruling class as percentages. The similarity middle from two points 'FRANCE' and 'FRENCH', for instance, is 40%. I'll start signifying likeness dossier as percentages, curved to the tightest number.

Typically, we clothe`t absolutely need to identify by what method corresponding strands are. We need to understand; that of a established of acknowledged strands are maximum just like a picked series. For example, that of the series 'Heard', 'Healthy', 'Help', 'Herded', 'Sealed' or 'Sold' is maximum just like the strand 'Healed'?

To answer the question, we just need to uncover the correspondence middle from two points 'Healed' and each of the additional agreements, and before rank the judgments orderly of these principles. Table displays the consequences concerning this case.

Table 1 Similarity

Word	**Similarity**
Sealed	80%
Healthy	55%
Heard	44%
Herded	40%
Help	25%
Sold	0%

IV. METHODOLOGY AND TOOLS

This part offered the technique for tackling the problems in a concise and technical manner. Also included are the architectural software and hardware, models employed, and data presented in the form of figures and tables to provide multiple abstraction perspectives of the solution.

The pertaining to syntax plan for question solving established DBpedia principles was submitted in this place study to handle the question of question handle, that was considered a critical and prerequisite treat for question solving arrangement research. As visualized in Figure 5, this method constituted of two fundamental processes: question transform and answer query convert and introduce C# Language.

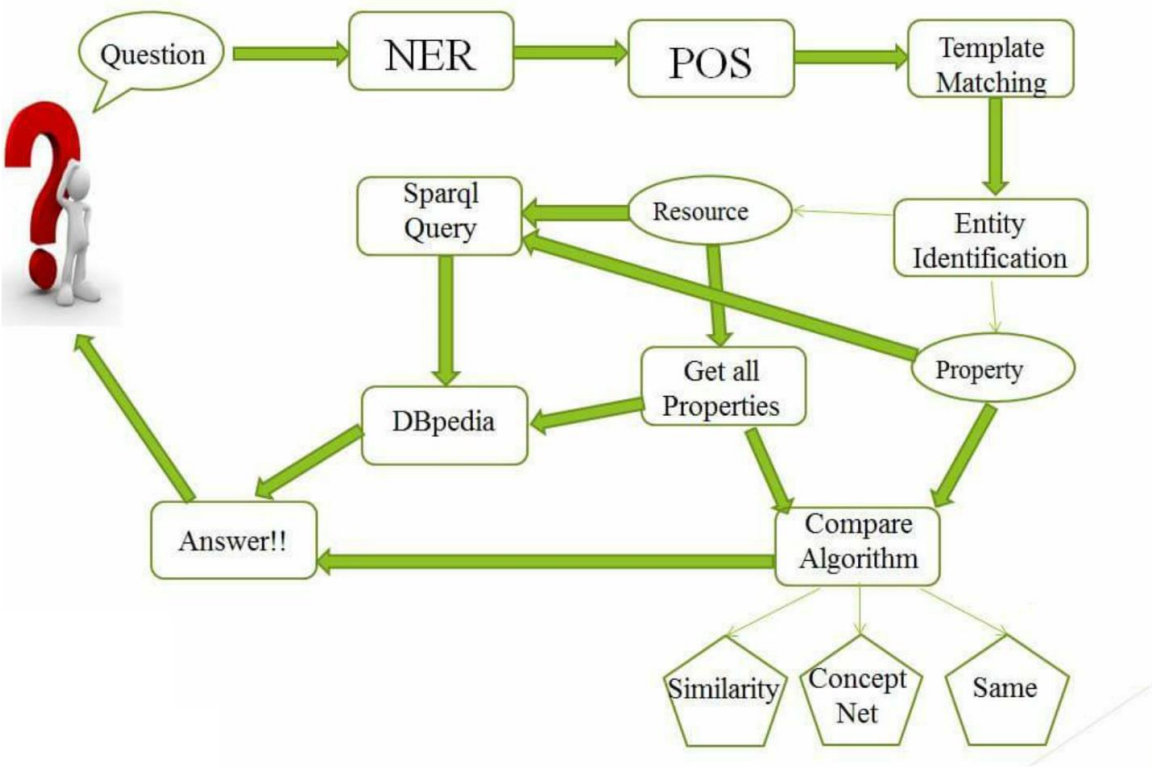

Figure 5 System Architecture

A. Question processing:

Step 1: Once the recommendation is captured from the consumer, it has expected treated. The question is shabby, kill the stop discussion, and is convinced into a standard layout avoid on NER.

The Named Entity Recognition (NER) the information bases appropriated to train the NE distillation invention have a considerable affect the process. Recent approaches to associate bodies to related dopes utilizing fine ontologies have existed bestowed, utilizing possessions to a degree DBpedia, Freebase, and YAGO. Attempts have existed fashioned to authorize patterns for disambiguating facts wholes utilizing a Uniform Resource Identifier (URI), apart from detecting a NE and allure kind. Because unrefined accents (as opposite to stiff or the study of computers) are essentially uncertain, disambiguation is individual of the basic issues in the study of computers, bestowing make even the extent of discussion-sense disambiguation (WSD). For example, contingent upon the circumstances, a article containing the phrase Washington can concern George Washington or Washington DC. People, institutions, and trades can all have a assortment of names and monikers. These systems usually expect clues in the encircling theme to help ponder and perfect the enigmatic term's engaged aim. As a result, a NE origin process requires analyzing recommendation material for chosen bodies, providing bureaucracy a type that is to say burden by their assurance score, and disambiguating bureaucracy accompanying upper class of URLs.

Step 2: Extract template tags by using Part of Speech Tag (POS), the process of giving each word in an input text a part-of-speech marker. Tokenization is normally done before, or as part of, the tagging process, because tags are frequently applied to punctuation: quotation marks, separating commas, etc., separating words from punctuation at the end of a phrase (period, question mark, etc.) punctuation derived from part-of-word punctuation (such as in abbreviations like e.g. and etc.) Considered as the tiniest of components with different meanings. Words are classified into numerous categories or sections of speech based on their use and functions. Noun, verb, pronoun, preposition, adjective, conjunction, interjection, adverb, and sometimes number, article, or determiner are all frequent English components of speech.

Step 3: Find the matching templates which already collected and stored. Then we are identifying the entities through determining recourse and property.

B. Answer query processing:

To get proper answer we have two ways: Once we identify the resource and the property from the question, then we can generate a SPARQL query in the specific format so that it can be hit on the DBpedia end point and then we can fetch the result.

If we didn't get any result, we will try with Compare algorithm, the steps of this algorithm are to reach to any synonyms word that can help us to reach the accurate answer.

Consider the following two examples:

Question 1: How tall is Cristiano Ronaldo?

Question 2: What is the Capital of Syria?

The source of the first question is "Cristiano Ronaldo". And "Syria" is the source of the second question. Once reserve and possessions are well established, gossip opportunity to rewrite the SPARQL query that will go for the answer from DBpedia. The query is define system and feature. To demonstrate this procedure, let's back to the previous question "What is the Capital of Syria?" as we seen previously, we got:

```
Select distinct * where {
<http://dbpedia.org/resource/ Syria >
<http://dbpedia.org/ontology/ Capital >?V.
OPTIONAL {filter (Lang (?V) ="EN")}
}
```

After extract all properties for the resource in question, we have three ways to find the result:

1- Same: The system tries to find same name of the property in our question and Dbpedia. Ex: (Capital == Capital).

2- String Similarity: The system tries to find name of the property in our question and nearest property name in Dbpedia Ex (Product == Products).

3-ConceptNet: If we can't find the same name of property in question and Dbpedia we can use Concept net to determine correct property. Ex (Wife == Spouse)

The Concept Net project aims to build a large semantic graph that describes general human knowledge and how it is expressed in natural language. Words and frequent phrases in every written human language are included in Concept Net's scope. It contains a vast amount of background information that any computer program working with natural language text should be aware of. The goal of Concept Net is to store both precise facts and the jumbled, inconsistent universe of common-sense information. It is critical to detect the informal relationships between ideas that are part of common knowledge and are often under-represented in other lexical resources in order to completely grasp concepts that exist in natural language writing.

Finally, the system interrogates the DBpedia Server to get response to retrieve the value of the property in DBpedia which is the answer of the question.

V. IMPLEMENTATION AND RESULTS

The evaluation is based on retrieving correct answer from DBpedia, It comprises of 20 questions over Dbpedia, and questions evaluated precision defined as follows:

$$\text{Precision} = \frac{(\text{Number of relevant answers retrieved})}{(\text{Total number of answers})}$$

Equation 3

Table 2 Sample of the question that tried on the system.

Number	Question	Relevant
1	How many children does Cameron have?	X
2	What is Muscular Dystrophy?	
3	Do beluga whales eat penguins?	
4	When was Dbpedia released?	X
5	What is largest city in Argentina?	X
6	Where is Detroit?	X
7	When was Albert born?	X
8	How many countries in Asia?	X
9	What language are spoken in Syria?	X
10	What is Russia's currency?	X
11	List all products of HTC?	X
12	What did Newton discover?	
13	Who is Cameron?	X
14	What is Russia's currency?	X
15	Why were the conquistadors important?	
16	What is the atmosphere Composition of mars?	X
17	Who is Danielle Steel wife?	X
18	Who invented papyrus?	
19	Where was Albert die?	X
20	What language are spoken in Syria?	X

$$\text{Precision} = \frac{15}{20} = 0.75\%$$

Equation 4

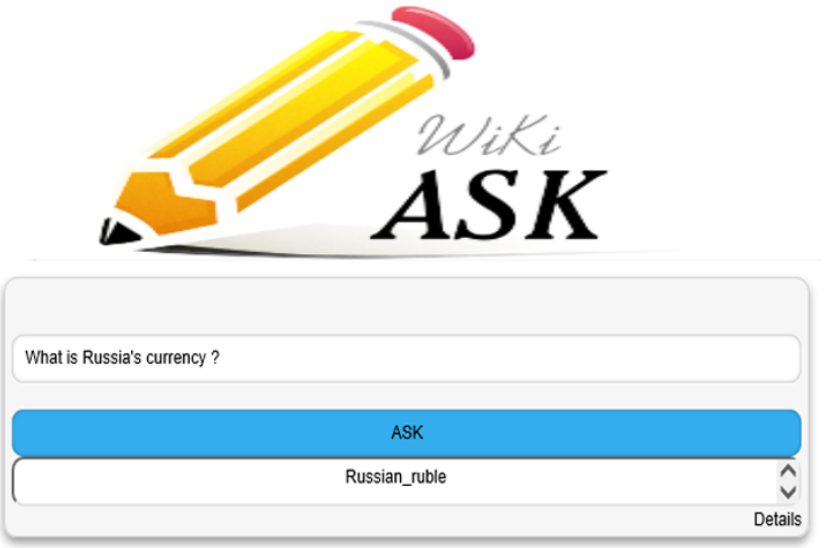

Figure 6 Home Screen

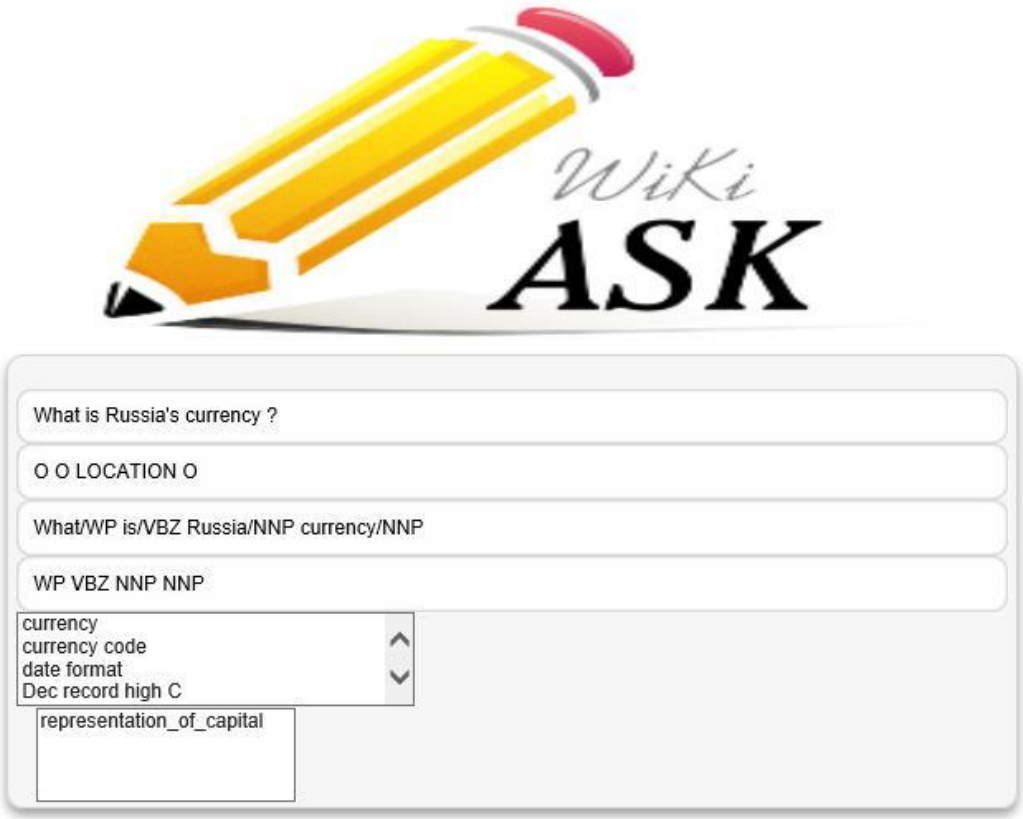

Figure 7 Details page

VI. CONCLUSIONS AND PROPOSAL

For the future study, we are planning to add additional lexical knowledge to the system (e.g. Wiktionary and lexical pattern libraries), integrating the system with more free bases to increase the accuracy by having access to more data from the network. Circulating system on several languages other than English and Increase the accuracy of the system. On a concluding note, we would like to present to you the “Intelligent Question Answering System” which gives you precise answers for a question asked a natural language. Abiding by the restrictions of closed domain QA system, our system is still performing far better compared to other QA system like QAKiS, which is also wiki framework-based system. Added advantage in using our system is that it tries to be as accurate and precise as possible by chopping down unwanted information.

VII. BIBLIOGRAPHY

BOOKS

Daniel Jurafsky and James H. M. **Speech and Language Processing**, University of Colorado at Boulder ,3rd Edition,2020.

Michael Workman. **Semantic Web.** Springer International Publishing Switzerland, 2016

Thomas R. G. **A Translation Approach to Portable Ontology Specifications.** Stanford University , 2009

Nicholas Gibbins and Nigel Shadbolt. **Resource Description Framework (RDF).** Southampton University , 2009

John Hebeler , Matthew Fisher , Ryan Blace and Andrew Perez-Lopez. **Semantic Web Programming**. Indianapolis,2009

Christian Bizer and Jens Lehmann**. DBpedia - A Crystallization Point for the Web of Data.** Leipzig University,Department of Computer Science, 2009

ARTICLES

[1] K. Sengloiluean, N. Arch-int, S. Arch-int\, T. Thongkrau,"A semantic approach for question answering using DBpedia and WordNet," International Joint Conference on Computer Science and Software Engineering (JCSSE), 2017.

[2] A. M. N. Allam and M. H. Haggag, "The question answering systems: A survey," International Journal of Research and Reviews in Information Sciences (IJRRIS), vol. 2, 2012.

[3] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question Answering Systems: Survey and Trends," Procedia Computer Science, vol. 73, pp. 366-375, 2015.

[4] M. H. Heie, E. W. D. Whittaker, and S. Furui, "Question answering using statistical language modelling," Computer Speech & Language, vol. 26, pp. 193-209, 2012.

[5] Jovita, Linda, A. Hartawan, and D. Suhartono, "Using Vector Space Model in Question Answering System," Procedia Computer Science, vol. 59, pp. 305-311, 2015/01/01 2015.

[6] A. Andrenucci, E. Sneiders,, "Automated Question Answering: Review of the Main Approaches," Stockholm University/ Royal Institute of Technolog,2005.

[7] C. Bizer, J. Lehmann, G Kobilarov, S. Auer, C. Becker, R. Cyganiak, S.Hellmannb ,"DBpedia - A Crystallization Point for r the Web of Data". Universitdt Leipzig,Department of Computer Science ,Leipzig,Berlin,2009.

[8] Lopez, V., Motta, E., Sabou, M., Fernandez, M.: PowerAqua: a multi-ontology based question answering system–v1. OpenKnowledge Deliverable D (2007)

[9] E. Cabrio, J. Cojan, A. P. Aprosio, B. Magnini, A. Lavelli, and F. Gandon, "QAKiS: an open domain QA system based on relational patterns," presented at the Proceedings of the 2012th International Conference on Posters & Demonstrations Track - Volume 914, Boston, USA, 2012.

Catherine Havasi, R. S. (2007). **ConceptNet : a Flexible, Multilingual Semantic Network for Common Sense Knowledge.**

ELECTRONIC SOURCES

McBride B. (2014, Feb 25). rdf-schema. Retrieved from W3C:

https://www.w3.org/TR/rdf-schema/

Daniel Fleischhacker, V. B. (2014, september 9). DBpedia Knowledge Base Version 2014 released. Retrieved from Data and Web Science Group

http://dws.informatik.uni-mannheim.de/en/news/singleview/detail/News/dbpedia-knowledge-base-version-2014-released/

DBpedia Lookup. (2015, April 29). Retrieved from dbpedia

https://en.wikipedia.org/wiki/DBpedia

Simon White.(1992). How to Strike a Match. Retrieved from catalysoft

http://www.catalysoft.com/articles/strikeamatch.html

RESUME

Name Surname **:** SARA MHD NASER JOUMA

EDUCATION :

- Bachelor : 2012-2016, Arab International University, Informatics Engineering, major of Artificial Intelligence
- **MA** **:** 2022, Istanbul Aydin University, Institute of Graduate Studies, Artificial Intelligence and Data since

PROFESSIONAL EXPERIENCE AND AWARDS:

- Admission and Registration representative and Being an academic advisor for new students at informatics and communication engineering faculty, Arab International University, July-August (2016)
- Mobile Developer, (Xamarin.Forms) , TMS Arabia, December (2016)
- Software Engineering and Quality Assurance , National Web Solutions Company - Mawaqaa Jordan , 15 March 2017 – 10 February 2020
- Translating Website from English to Arabic, freelancer
- Senior Software Engineering, Motorgy Company Kuwait, 15 February 2020 - until present
- Speak Arabic and English.
- I was one of the top students in Baccalaureate and I was in the first class more than one time through my university study

PUBLICATIONS FROM DISSERTATION, PRESENTATIONS AND PATENTS:

- AUTO QUESTION ANSWERING SYSTEM USING DBPEDIA (under processing).