A Mini-project Report

on

NATURAL LANGUAGE QUESTION ANSWER PAIR GENERATION

carried out as part of the course Information Retrieval (IT458)

Submitted by

Muthyam Satwik (16IT119)
Sai Manikanta Badugu (16IT209)

VI Sem B.Tech (IT)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

Department of Information Technology

National Institute of Technology Karnataka, Surathkal

Jan - May 2019

CERTIFICATE

This is to certify that the project entitled**" Natural Language Question Answering Pair Generation"**is a bonafide work carried out as part of the course **Information Retrieval (IT458)**, under my guidance by

1. Satwik Muthyam

2. Sai Manikanta Badugu

students of VI Sem B.Tech (IT) at the Department of Information Technology, National Institute of Technology Karnataka, Surathkal, during the academic semester of Jan-May2019 inpartial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology, at NITK Surathkal.

Place:

Date: Signature of the Instructor

DECLARATION

Wehereby declare that the project entitled "**Natural Language Question Answering Pair Generation**" submitted as part of the partial course requirements for the course Information Retrieval (IT458) for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal during the Jan - May2019 semester has been carried out by us. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Further, we declare that we will not share, re-submit or publish the code, idea, framework and/or any publication that may arise out of this work for academic or profit purposes without obtaining the prior written consent of the Course Instructor.

Name and signature of the Students:

1._____

2._____

Place:

Date:

Natural Language Question Answer Pair Generation

Satwik Muthyam
16IT119
Information Technology
muthyamsatwik@gmail.com

Sai Manikanta Badugu
16IT209
Information Technology
saimanikanta189@gmail.com

***Abstract*—Knowledge graphs such as Freebase that capture facts regarding entities and relationships between them have been used actively for responsive factoid questions. In this project, we have a tendency to explore the problem of mechanically generating question answer pairs from a given knowledge graph. The generated question answer (QA) pairs are often employed in many downstream applications. The generated QA pairs could be used for coaching higher QA systems. To come up with such QA pairs, we first extract a collection of keywords from entities and relationships expressed in a triple stored within the knowlegde graph. From each such set, we plan to use a set of keywords to generate a language question that has a unique answer. We then propose a sequence to sequence model RNN to form question answer pairs from the previously taken keyword tuples from the knowledge graph. Our model generates QA pairs with a BLEU score of 58.2.**

I. Introduction

The Knowledge Graph is a widely used data form by Google to improve the effectiveness of the search results by providing data collected from different sources .This information is shown to users in an infobox next to the search results.

From the knowledge graph we extract the information in the form of triples. We classify each node in the graph as either subject or object and the egde between the nodes the relation(predicate).

We take a triple having of a subject, predicate and object. Subject is defined over a domain while the object is defined over a range. The predicate gives the relationship between the subject and the range and each predicate may have multiple subject object pairs associated with it. In this model we consider only the case where the predicate is linked to a unique subject and object pair.

The keyword triples generated is used as a input to the model for Question generation and to give out a unique answer. For this, it is important to first select a subset of triples from which interesting and meaningful questions can be constructed. Further, even for an interesting triple, it may be possible to use only certain subsets of keywords to construct a meaningful question. We need to automatically identify the right set of keywords that should be used to form the question such that the answer also lies in the set. In this project we not only generate QA pairs but also propose a model to extract important keywords in the form of tuples from the knowledge graph given.

An example set of keywords constructed from the triple Capital(India, New Delhi)

Q: What is the capital of India?
A: New Delhi
Q: In which country is New Delhi?
A: India

We propose an RNN model which is trained over a set of questions taken from Wikianswers which consists of a vvariety of topics in it. We then used the keyword tuples generated in the previous step as input to the RNN model to generate a question with a unique answer associated with it.

II. Literature Survey

A number of papers have checked out the matter of generating vocabulary queries like WordNet(1990) and spacing similarity techniques. There are various works in automatic question generation from text. Many proposed ways based on syntax based ways that use the grammatic structure of sen-tences, determine key phrases and apply some transformation rules to make queries. Mannem et al. (2010) further used labeling for transformation rules. Cai et al. (2006) proposed associate XML nomenclature that used to manually produce question templates. This is sensitive to the performance of syntactical and linguistics parsing. Heilman and Smith(2010) used a rule based approach to remodel a declarative sentence into many candidate queries and rank them employing a supplying regression model. These approaches involve making templates manually and so need vast manual work and have low recall.

A problem that has been studied recently and is similar to our drawback of generating queries using data graph is that of generating queries from net queries. The motivation here is to mechanically generate queries from queries for community- based question respondent services such as WikiAnswers. The idea was implemmented by Lin and more developed by Zhao and Zheng. Each of those approaches were guide basedap-proaches where the templates were made to learn using a vast question corpus along side question logs. Dror et al.(2013) planned a learning to rank based technique to get grammati- cally correct and various queries from a given question where the candidate queries were measured by the model proposed by Zhao. These approaches use several question pairs to find out question templates and so have higher generalization performance compared to earlier methods wherever templates were learnt manually.

The most recent work is done by Seyler et al. (2015) who proposed a method to come up with language queries from data graphs given a subject of interest. They conjointly offer a technique to estimate problem of generated queries. The generation of question is done by manually created patterns and therefore is prescribed in application. In distinction we propose associate RNN based mostly technique to

findoutgeneration of language queries from a set of keywords.Themodel is trained using a dataset containing opendomainkeywords and questionpairs.

III. Motivation

Knowledge graph is a data structure that increasingly came into use to store data and extract information. Also generating natural language questions from keywords is a difficult task. This motivates us to generate question answer pairs from knowledge graph using deep learning model. The generated question answer pairs can be used in several applications like training better QA systems.

Generation of questions from a given data has always been an issue because of the syntax and semantics involved in the grammar. Many template based and syntax based models were proposed. But these models proved to be tidious and time consuming. So we came up with a deep learning which takes into account, the syntax and perform better in terms of speed and accuracy.

IV. ProblemStatement

Generation of Natural langauge question answer pairsfromagivenknowledgegraphusingaRNNmodel.

V. Objectives

- Toextract information from a given knowledgegraph.The data in the knowledge graph is present in thegraphas nodes and edges connecting them. Information istakenfromitbyformingtriplesusingtheedgesinthe graph.
- Togenerate natural language question answer pairswhichare unique. Each question generated is topic specificandthe set of keywords used in it are taken from aknowledgegraph.
- Totrain a RNN model which gives reliable questions to further develop a query system. The RNNmodelgenerates questions by taking the keywords generated from the knowledge graph. It is trained a datasetfromWikianswers.

VI. ProposedWork

Let KG be the information graph that contains information regarding numerous entities within the variety of triples.Atriple consists of an subject, a predicate and associated object. Subjects and objects are present in the form of nodesinthe KG, that might represent someone, a place, anabstractconception or a physical entity. Predicates are edgeswithinthe KG. They outline kind of relationship between thetopicand the object. The framework to get QA pairs consists two major modules. the primary module, QuestionKeywordsand Answer Extractor, is language independent andextractsneeded information about the entity E from the KG.The second module may be a language dependent RNN, primarily based Natural Language Question Generator. When fed with the knowledge extracted from the primary half it generates language QA pairs.

In order to retrieve data concerning the given entity E,wewant to initially determine the node n that represents theentityE within the KG. One way to identify node n is to leverage the label property. The next st[illegible] retrieve all the neighboursof n. Let m be a neighbour of n in K[illegible] by a predicate p.

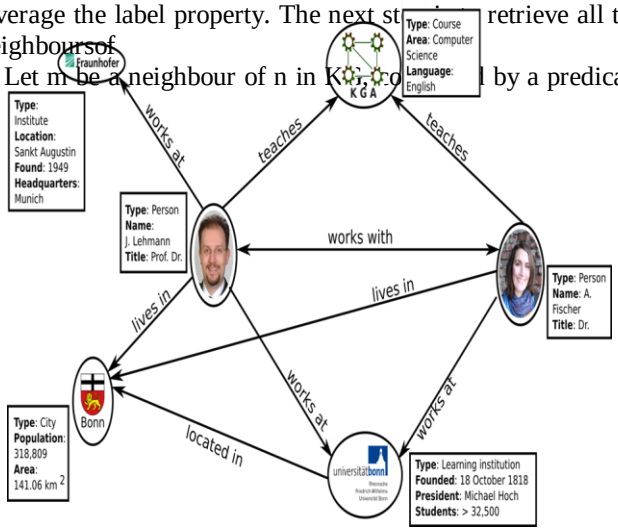

Fig. 1. Entity London in the Knowledge Graph

Figure 1 shows the knowledge graph with 6 entitieseachconnected to their respective objects through the predicates. For instance J.Lehmann is connected to Bonn throughthepredicate 'lives in'.Each of those neighbours are connected to the entity by a predicate. Given a predicate p, let sub(p) be the subject of p and obj(p) be the object of p. A predicate is usually outlined with a domain (subjecttype)and a range (object type) to produce higher linguistics. The domain and range defines the entity sorts that can beusedbecause the subject and object of the predicaterespectively.Let domain(p) and range(p) be the domain and range of p.Letsub(p),dom(p),p,obj(p),range(p) be the 5-tuple related toeachp.

Wethen describe QKA pairs are produced from 5-tuple.LetQk be the question keywords set and Ak be the solution tothequestion to be generated using Qk. (Qk, Ak) along cankinda QKApair.

A. *UniqueForwardRelation*

If p is unique for sub(p) in KG, then Q can include sub(p), p and range(p). AK will be obj(p). If p is not uniqueforsub(p), then there may be multiple potential answers tothegenerated question as well as obj(p), and therefore wedon'tgenerate such a QKpair.For the instance of keywordpairIndia,city,wecannotgenerateauniquequestion answerpair

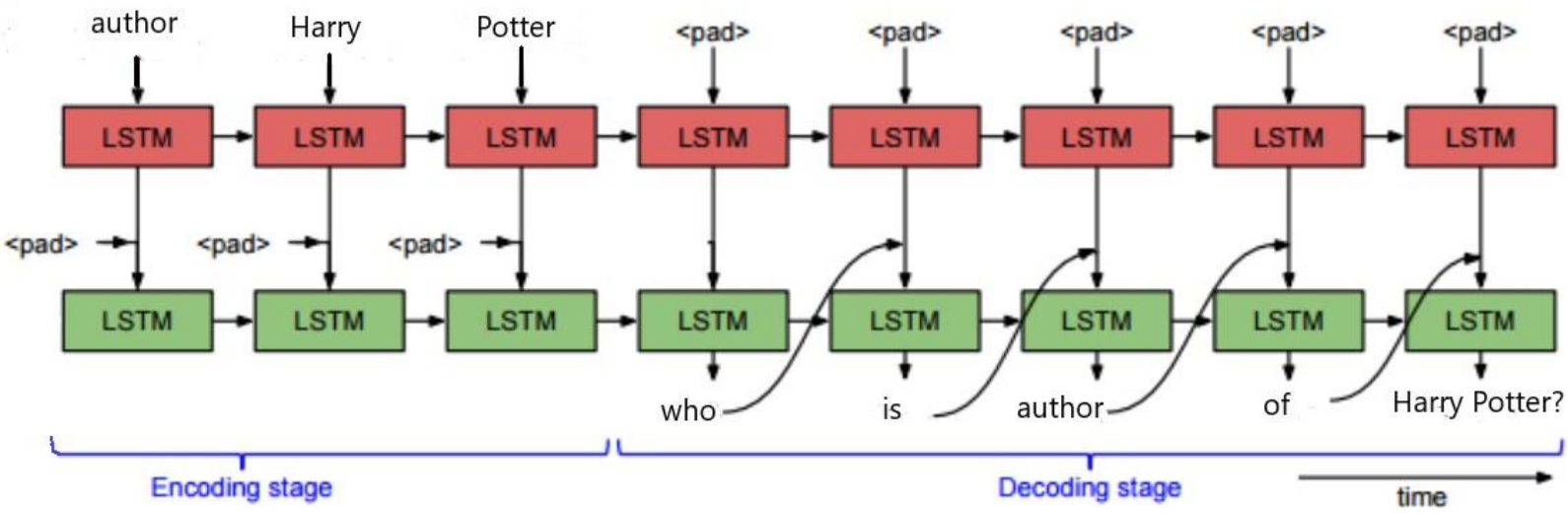

Fig. 2. RNN architecture

since India has many instances of city.

B. *Unique ReverseRelation*

If p is unique for obj(p) in KG, then Q can include obj(p), p and domain(p). For example Obama,President,USAcangenerateauniqueQApairwhentherever serelationisapplied.

In the previous section, we put forward the approachformaking question keywords and answer pairs. Then weproceedto model for generating natural language queries from agivensetofquestionkeywords.Wemakeuseofthekeywords,QK = qk1,qk2,.....,qkm, as an input sequence and the question, Q = q1,q2,....,qm, as the output sequence. This styleoftreating a collection of keywords as a sequence and notasa bag of words allows us to come up with totallydifferentsemanticallyvalidquestionsfromasimilarsetK.Forinstance, given the question keywords, QK = CEO, Google, wecangenerate 2 semantically valid queries by changing theorderof these words: (i) Who is the CEO of Google? and(ii)Does Google have a CEO?Wepropose a Natural Language Question Generation (NLQG) model that initially encodestheinput sequence using some distributed representation andthendecodes the output sequence from this encoded sequence. Specifically, we use a RNN based encoder and decoderforlanguage processtasks.

Let m be the length of input sequencea we converteachwordintovectoroffixedsizesuchthatit'ssizeislessthan n. The encoder maps each seaquence of xi to a fixedsizeencoding.Weuse RNN model to find value of hiusingthe formula: hi = (hi1, xi) where, hi ¡n is thehiddenrepresentation at position i.Wemake use of LSTM unitsasfortheimplementationofourmodel.Thedecoder calculates the probability of the output from the encoded vector.Thelength of the output sequence may or may not be the sameasthat of the input sequence. The output sequence is divided into conditional probabilities and the the sequence with thehighestprobability is given as the question for the dataprovided.p(qj—q¡j,hm)=t(qj1,gj,hm) where t is a non-linear function, that gives the probability of qj and gj is the decoder's hidden state. To train the RNN model we make use of the keywords generated from the open domain CQA website. During the runtime the keywords generated are fed to the RNN model which gives a permutation of possible questions and the one highest probability is choosen from them.

VII. ExperimentsandResults

A. *Dataset*

FortrainingtheK2Q-RNNmodelweneedasetofkeywordsand question pairs.Weuse a large collection of open-domain queriestakenfromWikiAnswers.Thisdatasethasaround20Mqueries.wehaverandomlychosen1Mqueriesfromthiscorpusfor coaching and 5k queries for testing (the most length ofaquestion was restricted to fifty words).Weextractkeywordsfrom the chosen queries by taking only Nouns,Verbsand Adjectives within the question. The components ofspeechtags were known usingNLTKparts of speech package.Weused the same order of the keywords taken from the question to maintain the meaning of the question.This sequenceofkeywords along side the initial question forms theinput-outputsequence used fortraining.

B. *Results*

The experiments conducted yielded results with a better performance than the existing models. We used BLEU score

Keywords	Expected Sentences	Predicted Sentences
plays jacob black twilight movies	who plays jacob black in the twilight movies	what plays in twilight movies
see outside paris	what to see outside of paris	where is outside of paris
christopher play batman returns	who does christopher walken play in abtman returns	who is batman returns
contribution maurice wilkins make dna	what contribution did maurice wilkins make to dna	what is the contribution of maurice wilkins

to evaluate the RNN model. The model gave results withaBLEU score of 58.2 . The loss while building the modelwasless than 3.0. The predicted results and the actual resultsareshown in the figure 2. Column 1 shows the keywords usedinthe training of the RNN model. Column 2 shows theactualquestions and column 3 shows the questions obtainedfromthe RNNmodel.

The results of the QA pair generator is shown in fig3.Column 1 shows the keywords fed to the RNN model.Thesekeywords are taken from the knowledge graph in theformo f t r i p l e s a n d t h e n c o n v e r t e d i n t o Q K A p a i r s . T h e secondcolumn shows the question generated for the givenkeywordsand the next column shows the answer for the generated question. Each question generated in the model has auniquecombination of Question andA n s w e r .

VIII. Conclusion

In this project we proposed a way of generating QApairsfor a given entity employing a knowlwedge graph.Wealsoworked on generating QA pairs using RNN model trainedwitha dataset of varied questions. This model worked wellgivingbetter results than the manual generation and templatebasedquestion generation which is tidious process.Wehavedoneevaluation for the model which showed that the modelworkedwell. This model can furthur be improved by adding features likestopwordremovalandadditionofadverbs.

IX. Timelineofproject

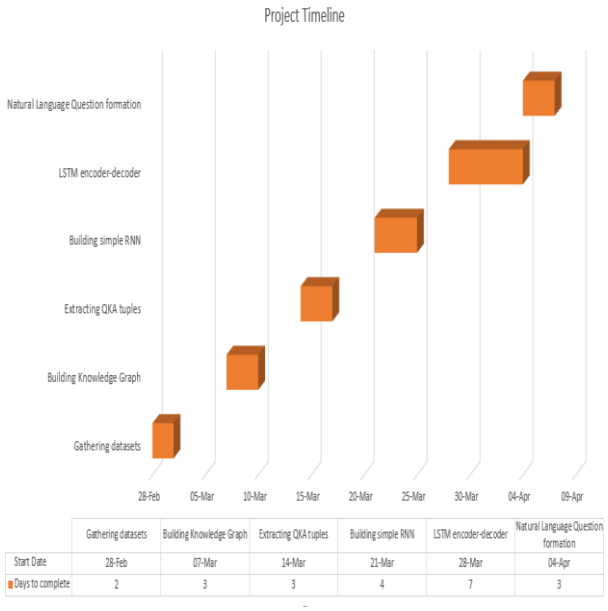

Fig. 3. Timeline of the project

Keywords	Predicted Sentences	Answer Extractor
Cuba treaties China	What are the treaties of Cuba	China
Bird exhibits behaviour	What bird exhibits in the?	Activities-Behaviour
China protests Countries	Who protests China?	India

X. IndividualContribution

16IT119 RNN Encoder/Decoder, KG Module Building,WordEmbeddings, Integration ofm o d u l e s

16IT209 Data Preprocessing QA Pair Generation,Simple RNN model

References

[1] Sathish Indurthi, Dinesh Raghu, Mitesh M. Khapra and Sachindra Joshi ”Generating Natural Language Question-Answer Pairsfroma Knowledge Graph Using a RNN Based Question Generation Mode”,Association for Computational Linguistics,2017.

[2] WeiguoZheng1, Jeffrey XuYu1,Lei Zou2, Hong Cheng, ”Question Answering Over Knowledge Graphs: Question UnderstandingViaTemplateDecomposition”,2018.