Generating Natural Language Question-Answer Pairs from a Knowledge Graph Using a RNN Based Question Generation Model

Samsung Electronics, DMC R&D

IBM Research India

Indian Institute of Technology Madras

Motivation

The generated question answer pairs can be used in downstream applications' performance.

In this paper, it proposed that triples from knowledge graphs can be used for automatically generating Question Answer pairs.

ex1.

Predicate	CEO
Subject	Sundar Pichai
Object	Google
Parent Predicate	designation
Domain	person
Range	organization
Keywords	CEO, designation, Sundar Pichai, person, Google, organization

Table 1: An example set of keywords constructed from the triple *CEO(Sundar Pichai, Google)*

Domain: subject type

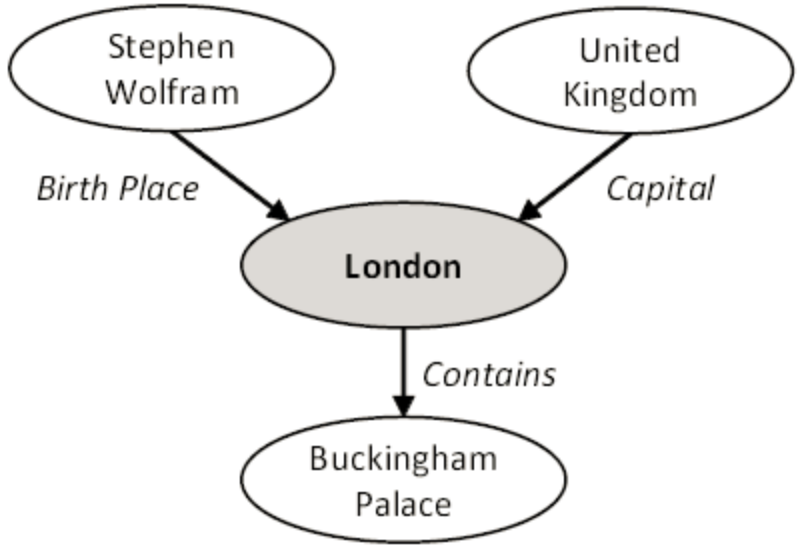
Range: object type

Contributions

1. It propsed a method for extracting triples and keywords from a knowledge graph for constructing question keywords and answer pairs.
2. It formulate the problem of generating natural language questions from keywords as seq-to-seq learning problem.
3. It trained the model using 1M questions from WikiAnswers ensuring that it is not tied to any specific knowledge graph.
4. QA system appended with the generated QA pairs get better performance.

Approach

Question Keywords and Answer Extractor



Retrieving all the neighbours of n.

	Column A	*Column B*	*Column C*
Subject Domain	United Kingdom Country	Stephen Wolfram Person	**London** Location
Predicate	Capital	Birth Place	Contains
Object Range	**London** City	**London** Location	Buckingham Palace Location

The rules to generate QKA pairs from 5-tuples.

Unique Forward Relaton:

If p_i is unique for $sub(p_i)$ in KG, then Q_k will include $sub(p_i)$,p_i and $range(p_i)$. A_k will be obj_{p_i}. 注：如果主语发生了这个谓语动作只能产生一个结果，那么取出。否则忽略。

Unique Reverse Relation:

If p_i is unique for $obj_{p_i)}$ in KG, then Q_k will include obj_{p_i}, p_i and $domain(p_i)$ 注: 如果宾语发生谓语动作只能产生一个结果，那么取出。否则忽略。

RNN based Natural Langeuage Queston Generator

Treat the keywords $QK = qk_1, ..., qk_m$ as an input sequence and the question, $Q = q_1, ..., q_l$ as the output sequence.
注：输入序列keywords是有序的，不同的顺序产生的问题应该是不同的。

encoder: LSTM1
decoder: LSTM2
so easy!!!

Training

Using a large collection of open-domain questions available from WikiAnswers dataset. Extracting keywords from the selected by retaining only Nouns, Verbs and Adjectives in question.
This sequence of keywords along with the original question forms one input-output sequence pair for training.

Inference

Beam search