Enriching Answers in Question Answering Systems using Linked Data

Rivindu Perera, Parma Nand, and Gisela Klette

School of Computer and Mathematical Sciences,
Auckland University of Technology, New Zealand
{rperera,pnand,gklette}@aut.ac.nz

Abstract. Linked Data has emerged as the most widely used and the most powerful knowledge source for Question Answering (QA). Although Question Answering using Linked Data (QALD) fills in many gaps in the traditional QA models, the answers are still presented as factoids. This research introduces an answer presentation model for QALD by employing Natural Language Generation (NLG) to generate natural language descriptions to present an informative answer. The proposed approach employs lexicalization, aggregation, and referring expression generation to build a human-like enriched answer utilizing the triples extracted from the entities mentioned in the question as well as the entities contained in the answer.

1 Introduction

Question Answering over Linked Data (QALD) offers new opportunities to traditional Question Answering (QA) systems by utilizing the massive Linked Data cloud as an information source. At its core, QALD transforms the natural language question to a SPARQL query and then execute it on a Linked Data resource to retrieve answers. These answers are then presented to the user as factoid answers without any further enhancements [1,2].

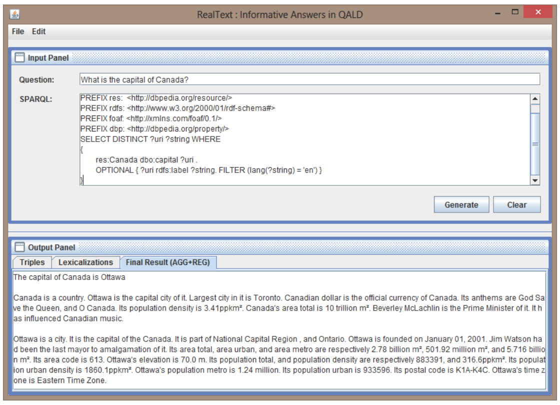
The RealText framework[1] described in this paper enhances the bare factoid answers by enriching them with more information and presenting them as natural text. An enriched answer is defined as an answer which provides a description of each of the entities contained in the question as well as in the answer to the question. Therefore, the enriched answer supports and validates the retrieved answer by providing background information more akin to a human generated answer. The RealText framework generates the description by using the triples related to the entity and application of a series of Natural Language Generation (NLG) techniques. In high level overview, these techniques can be categorized into lexicalization, aggregation, and Referring Expression Generation (REG), however each of these categories contain its own set of multiple subtasks to fine tune the final output.

[1] A video demonstration is available at https://vimeo.com/173608898

The rest of the paper presents an overview description of the framework features. Further details on some of the modules can be found in [3]. All features presented herein will be part of the demonstration.

2 Demonstration

The objective of the demonstration will be to present the complete RealText workflow from associating lexicalizing patterns to presenting an informative answer as natural text. The demonstration will use the RealText standalone application (for a screenshot see Fig. 1).



Question: What is the capital of Canada?

SPARQL:

```
PREFIX res:  <http://dbpedia.org/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbp: <http://dbpedia.org/property/>
SELECT DISTINCT ?uri ?string WHERE
{
      res:Canada dbo:capital ?uri .
      OPTIONAL { ?uri rdfs:label ?string. FILTER (lang(?string) = 'en') }
}
```

```
The capital of Canada is Ottawa

Canada is a country. Ottawa is the capital city of it. Largest city in it is Toronto. Canadian dollar is the official currency of Canada. Its anthems are God Sa
ve the Queen, and O Canada. Its population density is 3.41ppkm². Canada's area total is 10 trillion m². Beverley McLachlin is the Prime Minister of it. It h
as influenced Canadian music.

Ottawa is a city. It is the capital of the Canada. It is part of National Capital Region , and Ontario. Ottawa is founded on January 01, 2001. Jim Watson ha
d been the last mayor to amalgamation of it. Its area total, area urban, and area metro are respectively 2.78 billion m², 501.92 million m², and 5.716 billio
n m². Its area code is 613. Ottawa's elevation is 70.0 m. Its population total, and population density are respectively 883391, and 316.6ppkm². Its populat
ion urban density is 1860.1ppkm². Ottawa's population metro is 1.24 million. Its population urban is 933596. Its postal code is K1A-K4C. Ottawa's time z
one is Eastern Time Zone.
```

Fig. 1. A screenshot of the RealText desktop application

2.1 Datasets

For the demonstration we use the factoid questions extracted from the QALD-2 test dataset[2]. Since we work on the answer presentation (last step in QA) the input data comprised of question, SPARQL query, and as well as the extracted answer.

2.2 Workflow

The workflow comprises of three main modules; the *lexicalization module* which transforms the triples to natural language sentences, *Referring Expression Generation (REG) module* which assigns appropriate referring expressions to the mentions of the main entity, and *aggregation module* which aggregates individual sentences to form paragraphs. The final output contains the paragraphs as well as the answer in sentence form generated using our answer sentence generation framework [4].

[2] `http://qald.sebastianwalter.org/index.php?x=publications&q=2`

Lexicalization The objective of lexicalization module is to generate lexicalization patterns and associate them with triples. The framework is composed of four lexicalization pattern mining modules.

Occupational Metonym Patterns utilize the *-er* nominal based occupational metonyms to derive a predefined set of lexicalization patterns. For instance, a triple with occupational metonym, *director*, as the predicate and a movie as a subject. This triple can be lexicalized using a pattern such as $\langle$*S?, is directed by, O?* $\rangle_L$. We have developed a database which contains 33 of such patterns. These patterns are used to lexicalize a triple by matching the predicate and the core ontology class of the subject.

Context Free Grammar (CFG) Patterns uses the language generation capability of CFG and lexicalize the triples with past tense verb as a predicate. To be able to use CFG pattern, the verb (in predicate) should be identified as a verb having the frame, $NP \leftrightarrow VP \leftrightarrow NP$.

Relational Patterns use the unstructured text to derive patterns. We first pre-process text to resolve co-references and then extract relations ($\langle arg_1,\ rel,\ arg_2\ \rangle_R$) using OpenIE [5]. Each relation is then aligned with triples ($\langle subject,\ predicate,\ object\rangle_T$) to extract patterns. The alignment is calculated individually for subject and object alignment using Phrasal Overlap Measure (POM) and multiplied to get the final alignment score. Furthermore, we execute some realization steps using dependency parsing to resolve gender and grammar mismatches.

Property Patterns are predefined set of patterns which can lexicalize a given triple with specific predicate. For example, a pattern such as $\langle$*S?'s predicate, is, O?* $\rangle_L$ will be used to lexicalize triples with predicates, *population total*, *area total*, and *postal code*. There are five such patterns defined with their associated predicates from DBpedia [6].

We also carry out a realization phase after applying lexicalization patterns. The realization step corrects the syntactical errors of patterns such as a pattern does not match with the grammatical gender of the triple subject.

Table 1 shows some results from lexicalization modules where each triple is associated with a lexicalization pattern.

Aggregation The aggregation module first cluster the triples based on the subject. Then within each cluster we sub-cluster the triples based on rules. The triples within sub-clusters are then transformed to the natural language sentences using associated lexicalization patterns. However, at this level we do not substitute the subject expression (S?) of the sentence as it may need a referring expression in the generated paragraphs. Such referring expressions are resolved in the next phase.

Referring Expression Generation The referring expression generation module substitutes the subject expression with appropriate pronouns and entity names to emulate humans. In order to emulate this we change the referring expression after two consecutive usages.

Table 1. Sample set of triples, lexicalization patterns, and the pattern source. $S?$ and $O?$ denote subject and object respectively.

Triple	Pattern	Source	Score
$\langle$*Rubens Barrichello, birth place, Sao Paulo*$\rangle_T$	$\langle$*S?, was born in, O?* $\rangle_L$	Relational	0.8192
$\langle$*Rubens Barrichello, birth date, 1972-05-22*$\rangle_T$	$\langle$*S?, was born on, O?* $\rangle_L$	Relational	0.9028
$\langle$*Mount Everest, first ascent person, Edmund Hillary*$\rangle_T$	$\langle$*S?, was climbed by, O?* $\rangle_L$	Relational	0.4182
$\langle$*Captain America, creator, Joe Simon*$\rangle_T$	$\langle$*S?, was created by, O?* $\rangle_L$	Metonym	-
$\langle$*Lyndon B. Johnson, successor, Hubert Humphrey*$\rangle_T$	$\langle$*O?, succeeded, S?* $\rangle_L$	Metonym	-
$\langle$*London, population total, 8308369*$\rangle_T$	$\langle$*S?'s population total, is, O?* $\rangle_L$	Property	-
$\langle$*Canada, largest city, Toronto*$\rangle_T$	$\langle$*largest city in S?, is, O?* $\rangle_L$	Property	-
$\langle$*Socrates, influenced, Antisthenes*$\rangle_T$	$\langle$*S?, influenced, O?* $\rangle_L$	CFG	-
$\langle$*Intel, founded by, Robert Noyce*$\rangle_T$	$\langle$*S?, is founded by, O?* $\rangle_L$	CFG	-

3 Conclusion

This paper described the process of generating natural language descriptions for QALD. The approach is mainly inspired by the NLG where triple content is transformed to natural language paragraphs. In future we expect to extend the framework mainly focusing on the lexicalization pattern mining module. Furthermore, we will be looking at integration of this new approach to Intelligent Personal Assistant (IPA) to provide natural descriptions when presenting answers.

References

1. Perera, R., Nand, P.: Real text-cs - corpus based domain independent content selection model. In: ICTAI-2014. (2014) 599–606
2. Perera, R., Nand, P.: The role of linked data in content selection. In: PRICAI-2014. (2014) 573–586
3. Perera, R., Nand, P.: A Multi-strategy Approach for Lexicalizing Linked Open Data. CICLing (2015) 348–363
4. Perera, R., Nand, P.: Realtext-asg: A model to present answers utilizing the linguistic structure of source question. In: PACLIC-29, ACL Anthology (2015)
5. Mausam, Schmitz, M., Bart, R., Soderland, S., Etzioni, O.: Open language learning for information extraction. In: EMNLP, Jeju Island, ACL (jul 2012) 523–534
6. Bizer, C., Lehmann, J., Kobilarov, G.: DBpedia-A crystallization point for the Web of Data. Web Semantics **7**(3) (2009)