Assembling a large, cross-domain dataset for general data-to-text generation

Yale

Amrit Rau

Department of Computer Science, Yale University, New Haven, CT

LILY Lab

Background

A number of text-to-SQL systems exist to convert natural language queries to queries in SQL syntax. However, in order to close the loop on dialogue systems, it is important to generate natural language from the returned results. The majority of existing data-to-text literature focuses on the problem of generating multi-sentential natural realizations from an entire table; however, the task of generating a single-sentence realization from a single record is less well-covered.

Observations

Column ontology matters for fluent, semantically valid generation. Most realizations of individual records must take into account the ontology of the record's keys. Humans naturally infer ontologies when reading tables. The literature shows that incorporating the ontological structure of input data leads to measurably better realizations.

Semantic triples capture semantic meaning better than simple key-value records. Semantic (RDF) triplesets encode the relevant deep ontology by the graph structure of their relations.

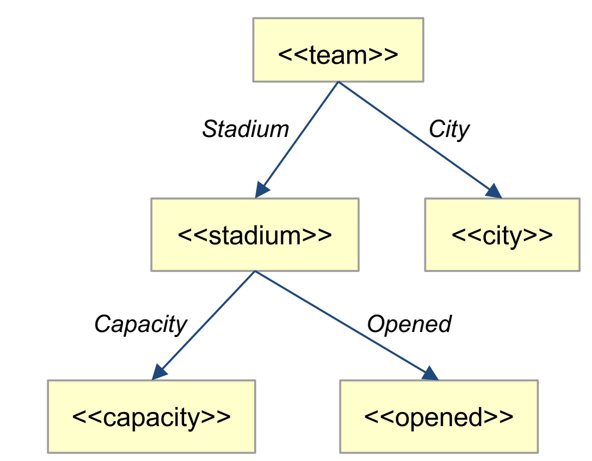
From a table record, its context, and its column ontology, a precisely semantically equivalent RDF tripleset can be constructed. See Figure 1 for an explanation of the relevant algorithm.

Task

Construct a large, cross-domain dataset of semantic triples aligned with natural language realizations. The semantic triples should be sourced from tables.

Figure 1. An simple algorithm to convert a table and its column ontology to a set of RDF triples.

Team	Stadium	Capacity	Opened	City
Amsterdam Admirals	Amsterdam ArenA	51,859	1996	Amsterdam, The Netherlands
Amsterdam Admirals	Olympisch Stadion	31,600	1928	Amsterdam, The Netherlands
Barcelona Dragons	Mini Estadi	15,276	1982	Barcelona, Spain



1. **Use ontology to construct subtables**

   Rule: for each non-leaf node, a subtable is generated containing that node's column and its immediate child nodes' columns.

Team	Stadium	City
Amsterdam Admirals	Amsterdam ArenA	Amsterdam
Amsterdam Admirals	Olympisch Stadion	Amsterdam
Barcelona Dragons	Mini Estadi	Barcelona

Stadium	Capacity	Opened
Amsterdam ArenA	51,859	1996
Olympisch Stadion	31,600	1928
Mini Estadi	15,276	1982

2. **Pivot each subtable on the parent node column**

   Each of the resulting rows in the pivoted subtables is an RDF triple.

Entity	Property	Value
Amsterdam Admirals	Stadium	Amsterdam ArenA
Amsterdam Admirals	Stadium	Olympisch Stadion
Amsterdam Admirals	City	Amsterdam
Barcelona Dragons	Stadium	Mini Estadi
Barcelona Dragons	City	Barcelona
Amsterdam ArenA	Capacity	51,859
Amsterdam ArenA	Opened	1996
Olympisch Stadion	Capacity	31,600
...	...	...

Figure 2. The workflow implemented to produce our dataset.

https://ppasupat.github.io/WikiTableQuestions/viewer/#203-760
URL http://en.wikipedia.org/wiki?action=render&curid=14213389&oldid=602141704
Title United States Ambassador to India

Name	State	Status	Title	Appointment	Credentials	Termination	Notes	
Henry F. Grady	California	Non-career appoir	Ambassador Extraordin	Apr 10, 1947	Jul 1, 1947	Left post, Jun 22, 194£	Accredited also to Nepal; resident at New Delhi.	
Loy W. Henderson	Colorado	Foreign Service ofi	Ambassador Extraordin	Jul 14, 1948	Nov 19, 1948	Reaccredited when Inc	Commissioned during a recess of the Senate; rec	
Chester Bowles	Connecticut	Non-career appoir	Ambassador Extraordin	Oct 10, 1951	Nov 1, 1951	Left post, Mar 23, 195:	Also accredited	Chester Bowles of Colorado was Aml
George V. Allen	North Carolina	Foreign Service ofi	Ambassador Extraordin	Mar 11, 1953	May 4, 1953	Left post, Nov 30, 195	Also accredited to Nepal; resident at New Delhi.	

1. User writes annotation on Google Sheet
2. Pipeline runs on server, retrieving Google Sheet
3. Annotations are aligned with key-value records
4. Heuristic is used to guess the subject column
5. Each key-value record is pivoted on the subject column to produce approximate triples
6. Triples are written to intermediate JSON format
7. Intermediate JSON file is converted to semantic triples using the algorithm shown in Fig. 1


```
<entry category="Entity" eid="Id1485" size="4">
  <modifiedtripleset>
    <mtriple>Chester Bowles | State | Connecticut</mtriple>
    <mtriple>Chester Bowles | Title | Ambassador Extraordinary and Plenipotentiary</mtriple>
    <mtriple>Chester Bowles | Appointment | 1951-10-10 00:00:00</mtriple>
    <mtriple>Chester Bowles | Termination of Mission | Left post, Mar 23, 1953</mtriple>
  </modifiedtripleset>
  <lex comment="good" lid="Id0">Chester Bowles of Colorado was Ambassador Extraordinary and Plenipotentiary to India from October 10, 1951 to March 23, 1953.</lex>
</entry>
```


Method

The dataset was assembled from tables from the WikiTableQuestions and WikiSQL corpora, both of which contain tables scraped from Wikipedia. The two corpora were combined and converted to spreadsheet form, then uploaded to Google Sheets. Subsequently, human annotators were instructed to generate natural language realizations from a subset of the keys for each row in the table. The annotators then highlighted the values in the row that were used in the generation. Once this was complete, the tables were pivoted to RDF form using the procedure described in Figure 2. Note that because we did not have each table's ontology, a heuristic was used to guess which column corresponded to the "subject" or principal entity of the table. This subject was used to construct an approximate ontology in which the subject was the sole parent and all other columns were children.

Findings

In this pilot study, **we generated a total of 1512 record-sentence** pairs out of a possible 151113. These lexicalizations **vary in topic, structure, and complexity.** We find that approximating the ontology **using a heuristic does not always produce semantically-correct triples**, specifically in tables with multi-column subjects or implicit subject. We also find that **table context** (e.g. title) **is often required to produce accurate semantic triples.** More work is needed to understand **how to best collect "gold" column ontologies** from human annotators or from other sources.

Acknowledgement

This work was performed under the guidance of Language, Information, and Learning at Yale (LILY), led by Dragomir Radev. I would like to acknowledge Rui Zhang, Jefferson Hsieh, Abhinand Sivaprasad, Aadit Vyas, Nazneen Rajani, and Dragomir Radev as collaborators on this work. I also thank Dragomir Radev for serving as a mentor and the formal advisor for this project.