DART: DAtabase Record to Text Generation

Aadit Vyas[1], Dragomir Radev PhD[1]

[1] Language Information and Learning Lab, Yale University, New Haven, CT

LILY Lab

Abstract

DART is a dataset for open domain semantic generation. It consists of annotated tuple-sentence pairs. We conducted a detailed analysis of DART and showed that it introduces new challenges compared to existing datasets, to which we have introduced better evaluation metrics. DART focuses on the task of generation from a single table record.

My main contribution to this paper was focused on converting annotations on WikiTableQuestions into a DART friendly format. More specifically, I converted row-level annotations into RDF triples for text generation.

Another task I worked on was also exploring Neural Wikipedian (Gardent et al., 2017a). Ultimately, we concluded that paper's model was not relevant to DART but that its data was potentially a good fit.

Materials and Methods

To build DART we used the following three methods: using existing datasets as-is (e.g. WebNLG (Gardent et al., 2017a) and E2E (Dušek et al., 2018)), converting existing datasets such as COSQL (Yu et al., 2019a), and manually annotating from tables in Spider (Yu et al., 2018) and WikiTableQuestions (Pasupat and Liang, 2015).

As a proof-of-concept, members of LILY manually annotated rows of tables on a Google Sheet. Collectively, we annotated about 2,000 rows and generated 16,931 unique triples (Table 2). This was an important step for validating the dataset. To gather more annotations, Amazon MTurk will be a great platform through which to solicit many more sentences for row-level data.

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 appoin	Ambassador Extraordina	Apr 10, 1947	Jul 1, 1947	Left post, Jun 22, 1948	Accredited also	
Loy W. Henderson	Colorado	Foreign Service off	Ambassador Extraordina	Jul 14, 1948	Nov 19, 1948	Reaccredited when Ind	Commissioned	
Chester Bowles	Connecticut	Non-career appoin	Ambassador Extraordina	Oct 10, 1951	Nov 1, 1951	Left post, Mar 23, 1953	Also accredited	Chester Bowles c
George V. Allen	North Carolina	Foreign Service off	Ambassador Extraordina	Mar 11, 1953	May 4, 1953	Left post, Nov 30, 1954	Also accredited	
John Sherman Coop	Kentucky	Non-career appoin	Ambassador Extraordina	Feb 4, 1955	Apr 9, 1955	Left post, Apr 23, 1956	Also accredited	
Ellsworth Bunker	Vermont	Non-career appoin	Ambassador Extraordina	Nov 28, 1956	Mar 4, 1957	Left India, Mar 23, 196	Also accredited	Ellsworth Bunker
John Kenneth Galbra	Massachusetts	Non-career appoin	Ambassador Extraordina	Mar 29, 1961	Apr 18, 1961	Left post, Jul 12, 1963		
Chester Bowles	Connecticut	Non-career appoin	Ambassador Extraordina	May 3, 1963	Jul 19, 1963	Left post, Apr 21, 1969		Chester Bowles c
Kenneth B. Keating	New York	Non-career appoin	Ambassador Extraordina	May 1, 1969	Jul 2, 1969	Left post, Jul 26, 1972		Kenneth B. Keati
Daniel P. Moynihan	New York	Non-career appoin	Ambassador Extraordina	Feb 8, 1973	Feb 28, 1973	Left post, Jan 7, 1975		
William B. Saxbe	Ohio	Non-career appoin	Ambassador Extraordina	Feb 3, 1975	Mar 8, 1975	Left post, Nov 20, 1976		
Robert F. Goheen	New Jersey	Non-career appoin	Ambassador Extraordina	Apr 26, 1977	May 26, 1977	Left post, Dec 10, 1980		

Table 1. Example of manual annotation on WikiTableQuestions


```
"dataset": "WikiTableQuestions",
"id": "203-760",
"path": "WikiTableQuestions/output/203-760.json",
"records": [
    {
        "record": {
            "Name": "Henry F. Grady",
            "State": "California",
            "Status": "Non-career appointee",
            "Title": "Ambassador Extraordinary and Plenipotentiary",
            "Appointment": "Apr 10, 1947",
            "Credentials Presented": "Jul 1, 1947",
            "Termination of Mission": "Left post, Jun 22, 1948",
            "Notes": "Accredited also to Nepal; resident at New Delhi."
        },
        "annotations": []
    },
    {
        "record": {
            "Name": "Loy W. Henderson",
            "State": "Colorado",
            "Status": "Foreign Service officer",
            "Title": "Ambassador Extraordinary and Plenipotentiary",
            "Appointment": "Jul 14, 1948",
            "Credentials Presented": "Nov 19, 1948",
            "Termination of Mission": "Reaccredited when India became a republic; presented
            "Notes": "Commissioned during a recess of the Senate; recommissioned after conf
        },
        "annotations": []
    },
    {
        "record": {
            "Name": "Chester Bowles",
            "State": "Connecticut",
            "Status": "Non-career appointee",
            "Title": "Ambassador Extraordinary and Plenipotentiary",
            "Appointment": "Oct 10, 1951",
            "Credentials Presented": "Nov 1, 1951",
            "Termination of Mission": "Left post, Mar 23, 1953",
            "Notes": "Also accredited to Nepal; resident at New Delhi."
        },
        "annotations": []
    },
    {
        "record": {
            "Name": "George V. Allen",
            "State": "North Carolina",
            "Status": "Foreign Service officer",
            "Title": "Ambassador Extraordinary and Plenipotentiary",
            "Appointment": "Mar 11, 1953",
            "Credentials Presented": "May 4, 1953",
            "Termination of Mission": "Left post, Nov 30, 1954",
            "Notes": "Also accredited to Nepal; resident at New Delhi."
        },
        "annotations": []
    },
```


Figure 1. Example of JSON intermediary table


```
<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 Plenipot
</entry>
```


Figure 2. Example of an RDF triple produced from JSON intermediary
Image from Amrit Rau

Dataset Name	Number of unique triplesets	Number of unique sentences	Potential number of sentences
WikiTableQuestions	16931	1512	16931

Table 2. Summary of WikiTableQuestions proportion in DART

	BLEU	METEOR	TER
GTR-LSTM	54.00	0.37	0.45
GCN-EC	55.90	0.39	0.41
GRU	56.09	0.42	0.39
Transformer	56.28	0.42	0.39
Step-by-Step	53.30	0.44	0.47
PlanEnc	64.42	0.45	0.33
DualEnc	63.45	0.46	0.34

Table 3. Results on salient event identification
Image from Rui Zhang

DART Dataset

Automatically generating textual description from structured input is critical to improving accessibility of database to lay users. As data-to-text generation garnered increasing attention in recent years, effort in both dataset and model development have been driving progress in this field.

DART focuses on providing a large, open domain corpus, with each input being a semantic tuple from a database record, annotated with sentence description that covers all facts in the tuple and attends to information from the table schema and table.

For WikiTableQuestions, manually annotated tables (Table 1) were converted to an intermediary JSON format (Figure 1) which is then converted to an RDF triple (Figure 2) that can be fed to a natural language generation model.

Conclusions

Early results using WebNLG are shown on Table 3. More research is being done to improve the sentence generation from DART and to maximize the utility of incorporating human annotations into the language generation model.

In all, DART presents a promising step forward in data-to-text generation by providing a robust dataset that is open domain and accessible for language models to leverage.

Using the RDF triples to make DART compatible with existing models shows how DART can be a generalizable platform through which row-level data from any domain can be used to generate natural language output.

Acknowledgements

Thank you to Dragomir Radev, Nazneen Rajani, Rui Zhang, Amrit Rau, and Abhi Sivaprasad for their guidance and support on this project.