Yale

# Natural Language Generation for Tabular Data

Abhi Sivaprasad

Department of Computer Science, Yale University

LILY Lab

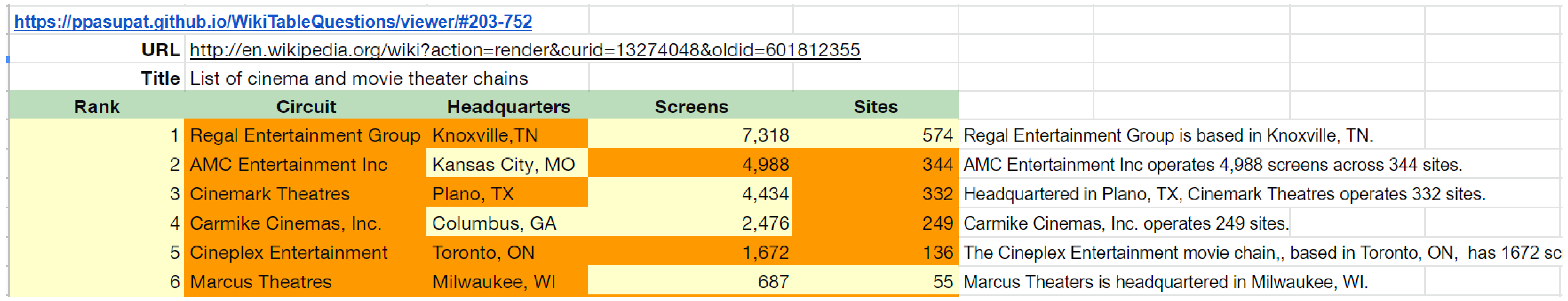
## Introduction

There is a tremendous amount of structured information found in open source knowledge bases. However, these troves of data are hard to navigate–they may require complex queries to retrieve appropriate information and examination to understand the returned records of data. The gold standard would be a dialogue agent which can take in queries and output answers in natural language. This breaks down the problem into two parts: converting a natural language query to a database query language (SQL) and converting the returned records into a natural language description. This work focuses on the latter problem by constructing a multi-domain dataset of tabular data with human annotated lexicalizations. In addition, this paper proposes a model to convert the tabular data into natural language descriptions.

## Materials and Methods

DART was built by accumulating tables from several data sources including E2E, WikiSQL, COSQL, Spider, and WikiTableQuestions. For each record in a table, annotators selected a random subset of columns, indicated by shading the cells orange (Figure 1). Then, they write a natural language sentence which describes the information in the highlighted cells. The data is extracted as RDF triples with attached lexicalizations. Additionally, for extra context for the model, the entities are piped into named entity recognition.

The input to the models is shown in Figure 3. A set of generalized triples, which replace exact instances with the general category, is input to the model. The human translation is similarly delexicalized by finding objects in the string and replacing them with categories. As DART is still under construction, the models have been trained and tested on WebNLG. Several variants of seq2seq with attention architectures have been tried including bidirectional, cnn, transformer, and stacked encoders.



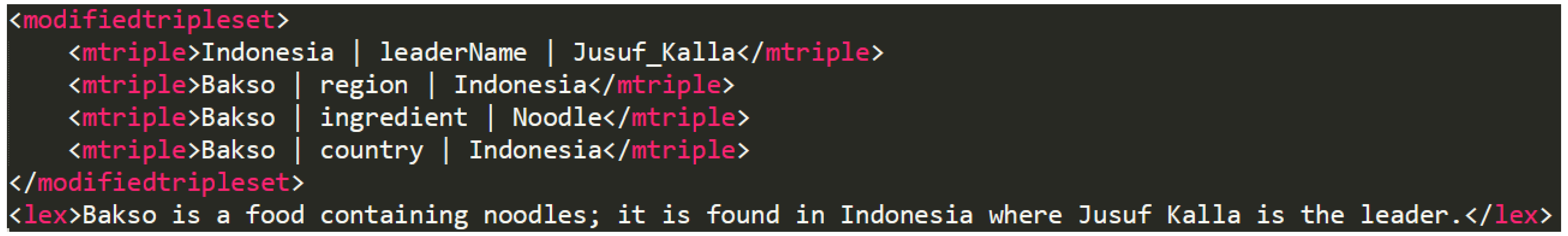
https://ppasupat.github.io/WikiTableQuestions/viewer/#203-752

**URL** http://en.wikipedia.org/wiki?action=render&curid=13274048&oldid=601812355

**Title** List of cinema and movie theater chains

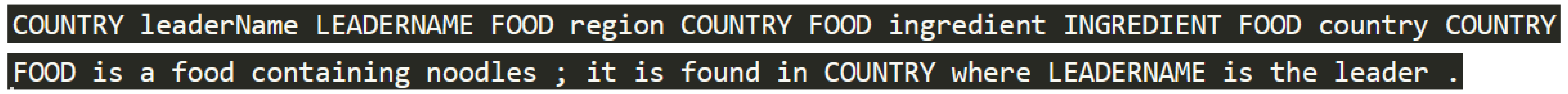
| Rank | Circuit | Headquarters | Screens | Sites | |
|---|---|---|---|---|---|
| 1 | Regal Entertainment Group | Knoxville,TN | 7,318 | 574 | Regal Entertainment Group is based in Knoxville, TN. |
| 2 | AMC Entertainment Inc | Kansas City, MO | 4,988 | 344 | AMC Entertainment Inc operates 4,988 screens across 344 sites. |
| 3 | Cinemark Theatres | Plano, TX | 4,434 | 332 | Headquartered in Plano, TX, Cinemark Theatres operates 332 sites. |
| 4 | Carmike Cinemas, Inc. | Columbus, GA | 2,476 | 249 | Carmike Cinemas, Inc. operates 249 sites. |
| 5 | Cineplex Entertainment | Toronto, ON | 1,672 | 136 | The Cineplex Entertainment movie chain,, based in Toronto, ON,  has 1672 sc |
| 6 | Marcus Theatres | Milwaukee, WI | 687 | 55 | Marcus Theaters is headquartered in Milwaukee, WI. |

Figure 1. A table from WikiTableQuestions annotated. The columns used for each record are shown in orange. Lexicalizations are on the right.



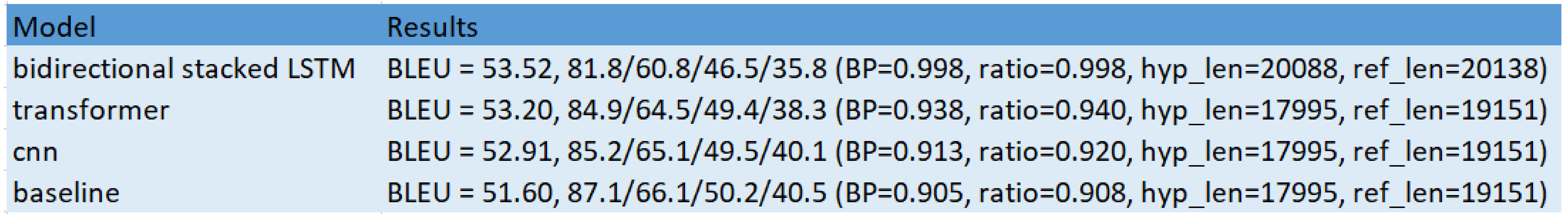
```
<modifiedtripleset>
    <mtriple>Indonesia | leaderName | Jusuf_Kalla</mtriple>
    <mtriple>Bakso | region | Indonesia</mtriple>
    <mtriple>Bakso | ingredient | Noodle</mtriple>
    <mtriple>Bakso | country | Indonesia</mtriple>
</modifiedtripleset>
<lex>Bakso is a food containing noodles; it is found in Indonesia where Jusuf Kalla is the leader.</lex>
```

Figure 2. An example of a processed set of RDF triples with corresponding natural language description



```
COUNTRY leaderName LEADERNAME FOOD region COUNTRY FOOD ingredient INGREDIENT FOOD country COUNTRY
FOOD is a food containing noodles ; it is found in COUNTRY where LEADERNAME is the leader .
```

Figure 3. Top is the set of generalized triples from the set in Figure 2. Bottom is the delexicalized natural language description



| Model | Results |
|---|---|
| bidirectional stacked LSTM | BLEU = 53.52, 81.8/60.8/46.5/35.8 (BP=0.998, ratio=0.998, hyp_len=20088, ref_len=20138) |
| transformer | BLEU = 53.20, 84.9/64.5/49.4/38.3 (BP=0.938, ratio=0.940, hyp_len=17995, ref_len=19151) |
| cnn | BLEU = 52.91, 85.2/65.1/49.5/40.1 (BP=0.913, ratio=0.920, hyp_len=17995, ref_len=19151) |
| baseline | BLEU = 51.60, 87.1/66.1/50.2/40.5 (BP=0.905, ratio=0.908, hyp_len=17995, ref_len=19151) |

Figure 4. Selection of models. Baseline is the reproduced WebNLG baseline provided in their 2017 competition.

## Results

The resulting DART database currently has 151113 (data, text) pairs of which 1512 have been annotated in the pilot. We have found that it's difficult to automatically extract ontologies from a set of columns so human annotators must mark them for each table. Often, this requires the inclusion of the table title which may be the subject for the table.

The performance of the models are shown in Figure 4. These models are preliminary and don't include copy mechanisms or pre-trained word embeddings. The baseline was provided by WebNLG, and has been surpassed by several of our seq2seq models. The best performing model was a bidirectional stacked LSTM with two layers, outperforming the baseline by 2 BLEU points.

## Limitations & Future Work

Two problems with the data are limited ontology and many relations with few examples. To address this, we are exploring Dbpedia, an RDF knowledge base extracted from Wikipedia. From DBPedia, we can gain precise control over the full ontology and can decide the count of triples for each relation. For the current entries in DART, annotators will need to mark the pivot columns to extract ontology.

To improve model architectures, BERT ensembles are the new world order. This requires extensive tuning and experimentation. In addition, work has been done by Marcheggiani et. al. which uses a graph based convolutional network encoder to exploit the structure of RDF structure. Similar encoders can be explored which also integrate the BERT architecture which has proven to be so successful.

## Acknowledgement

I would like to thank Dragomir Radev for being my formal advisor for this project. In addition, I would like to thank Rui Zhang, Aadit Vyas, Amrit Rau, and Nazneen Rajani for their contributions.