From Tables to Knowledge: Recent Advances in Table Understanding (Part III)

Neural Representation Learning on Tables

Huan Sun
The Ohio State University
Outline: Neural Representation Learning on Tables

• Background
Outline: Neural Representation Learning on Tables

• Background

• Representative Methods (pre-training objectives, table types, tasks)
  • Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR’19 (short); University of Stavanger]
  • TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB’21; OSU & Google]
  • TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL’21; Sony Co. & Adobe Research & UMass Amherst]
  • TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD’21; MSR & CMU & PKU]
Outline: Neural Representation Learning on Tables

• Background

• Representative Methods
  • Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval  
    [Deng et al., SIGIR’19 (short); University of Stavanger]
  • TURL: Table Understanding through Representation Learning  
    [Deng et al., VLDB’21; OSU & Google]
  • TABBIE: Pretrained Representations of Tabular Data  
    [Iida et al., NAACL’21; Sony Co. & Adobe Research & UMass Amherst]
  • TUTA: Tree-based Transformers for Generally Structured Table Pre-training  
    [Wang et al., SIGKDD’21; MSR & CMU & PKU]

• Summary

• Resources
Outline: Neural Representation Learning on Tables

• Background

• Representative Methods
  • Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR’19 (short); University of Stavanger]
  • TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB’21; OSU & Google]
  • TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL’21; Sony Co. & Adobe Research & UMass Amherst]
  • TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD’21; MSR & CMU & PKU]

• Summary

• Resources
Background: Representation Learning on Text

- **Word2Vec [Mikolov et al., 2013]**

CBOW

- Input: $w(t-2), w(t-1), w(t+1), w(t+2)$
- Projection: SUM
- Output: $w(t)$

Skip-gram

- Input: $w(t)$
- Projection: $w(t-1), w(t-2), w(t+1), w(t+2)$
- Output: $w(t)$
Background: Contextualized Representation Learning and Pre-training

Figure credit: https://github.com/thunlp/PLMpapers
Background: Transformer [Vaswani et al., 2017]

\[ z_i = \sum_{j=1}^{n} \alpha_{ij} (x_j W^V) \]

\[ \alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}} \]

\[ e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_z}} \]

Figure credit: https://www.arxiv-vanity.com/papers/1908.04211
Background: BERT [Devlin et al., 2019]

1. Self-supervised objectives: Masked language model (MLM) and Next Sentence Prediction (NSP)
2. Encoder: Transformer encoder [Vaswani et al., 2017]
3. Various downstream tasks such as Question Answering, Named Entity Recognition, etc.
Background: Previous tutorials

• Tutorial on Web Table Extraction, Retrieval and Augmentation
  • Shuo Zhang and Krisztian Balog, SIGIR 2019
  • https://iai-group.github.io/webtables-tutorial/

• Tutorial on Table Extraction and Understanding for Scientific and Enterprise Applications
  • Doug Burdick, Alexandre V Evfimievski, Nancy Wang, Yannis Katsis, Marina Danilevsky, VLDB 2020
  • https://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=10534

For more about definitions of various tasks and other approaches that are not based on table representation learning
Neural Representation Learning on Tables

• Different from “Using Deep Learning for Table-based Tasks”

• General-purpose, not task-specific representations of table elements

• “Unsupervised” representation learning
  • No human annotation
  • Self-supervised data
  • Self-supervised tasks
Outline: Neural Representation Learning on Tables

• Background

• Representative Methods
  • Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR’19 (short); University of Stavanger]
  • TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB’21; OSU & Google]
  • TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL’21; Sony Co. & Adobe Research & UMass Amherst]
  • TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD’21; MSR & CMU & PKU]

• Summary

• Resources
Outline: Neural Representation Learning on Tables

• Background

• Representative Methods
  • Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR’19 (short); University of Stavanger]
  • TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB’21; OSU & Google]
  • TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL’21; Sony Co. & Adobe Research & UMass Amherst]
  • TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD’21; MSR & CMU & PKU]

• Summary

• Resources
Table2Vec: Neural Word and Entity Embeddings

- Backbone algorithm: Word2Vec (Skip-Gram) [Mikolov et al., NIPS'13]

\[
\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)
\]

Figure credit: https://d2l.ai/chapter_natural-language-processing-pretraining/word2vec.html

[Chen et al., SIGIR’19 (short)]
Table2Vec: Neural Word and Entity Embeddings

• Backbone algorithm: Word2Vec [Mikolov et al., NIPS’13]

• Four variants, depending on what info in a table is modeled

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Semantic repr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table2VecW</td>
<td>all table data</td>
<td>word embeddings</td>
</tr>
<tr>
<td>Table2VecH</td>
<td>table headings</td>
<td>heading embeddings</td>
</tr>
<tr>
<td>Table2VecE</td>
<td>all entities</td>
<td>entity embeddings</td>
</tr>
<tr>
<td>Table2VecE*</td>
<td>core column entities</td>
<td>entity embeddings</td>
</tr>
</tbody>
</table>

[Deng et al., SIGIR’19 (short)]
Table2Vec: Neural Word and Entity Embeddings

• Backbone algorithm: Word2Vec [Mikolov et al., NIPS'13]

• Four variants

• Tasks
  • Row population
  • Column population
  • Table retrieval

Some results will be discussed later together with more advanced methods’

[Deng et al., SIGIR'19 (short)]
Outline: Neural Representation Learning on Tables

• Background

• Representative Methods
  • Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR'19 (short); University of Stavanger]
  • TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB’21; OSU & Google]
  • TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL’21; Sony Co. & Adobe Research & UMass Amherst]
  • TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD’21; MSR & CMU & PKU]

• Summary

• Resources
TURL: Table Understanding through Representation Learning

- Goal: (1) To learn deep *contextualized* representations of table elements;

![National Film Award for Best Direction](image)

From Wikipedia, the free encyclopedia

<table>
<thead>
<tr>
<th>Year</th>
<th>Recipient</th>
<th>Film</th>
<th>Language</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>1968 (16th)</td>
<td>Satyajit Ray</td>
<td><em>Goopy Gyne Bagha Byne</em></td>
<td>Bengali</td>
<td>[14]</td>
</tr>
<tr>
<td>1970 (18th)</td>
<td>Satyajit Ray</td>
<td><em>Pratidwandi</em></td>
<td>Bengali</td>
<td>[16]</td>
</tr>
</tbody>
</table>

[Deng et al., VLDB’21]
TURL: Table Understanding through Representation Learning

- Goal: (1) To learn deep *contextualized* representations of table elements;
  - To capture row-and-column structure
  - To capture factual knowledge in relational tables

From Wikipedia, the free encyclopedia

Winners [edit]

<table>
<thead>
<tr>
<th>Year</th>
<th>Recipient</th>
<th>Film</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>1967 (15th)</td>
<td>Satyajit Ray</td>
<td>Chiriyakhana</td>
<td>Bengali</td>
</tr>
<tr>
<td>1968 (16th)</td>
<td>Satyajit Ray</td>
<td>Goopy Gyne Bagha Byne</td>
<td>Bengali</td>
</tr>
<tr>
<td>1969 (17th)</td>
<td>Mrinal Sen</td>
<td>Bhuvan Shome</td>
<td>Hindi</td>
</tr>
<tr>
<td>1970 (18th)</td>
<td>Satyajit Ray</td>
<td>Pratidwandi</td>
<td>Bengali</td>
</tr>
</tbody>
</table>

page title & topic entity
section title
caption
headers
entity
object columns

[Deng et al., VLDB’21]
TURL: Table Understanding through Representation Learning

- Goal: (I) To learn deep *contextualized* representations of table elements;
  (II) Pre-training / Fine-tuning paradigm for table-based tasks to reduce feature engineering effort;

[Deng et al., VLDB’21]
TURL: Table Understanding through Representation Learning

• Goal: (I) To learn deep *contextualized* representations of table elements;
  (II) Pre-training / Fine-tuning paradigm for table-based tasks to reduce feature engineering effort;
  (III) New datasets for a series of table understanding tasks

[Deng et al., VLDB’21]
TURL: Table Understanding through Representation Learning

- Model: Transformer [Vaswani et al., NeurIPS 2017] with masked self-attention

[Deng et al., VLDB’21]
TURL: Table Understanding through Representation Learning

- **Model:** Transformer [Vaswani et al., NeurIPS 2017] with masked self-attention

[Image of a diagram showing a graph representation of a table]

- **Masked self-attention:** A table is treated as a graph and each component can only aggregate information from its neighbors

---

[From Tables to Knowledge (KDD’21): Jay Pujara, Pedro Szekely, Huan Sun, Muhao Chen]
**TURL: Table Understanding through Representation Learning**

- **Model:** Transformer [Vaswani et al., NeurIPS 2017] with masked self-attention

---

[Image of the TURL model diagram with tokens, entities, and Transformer layers, including Pre-training masked language model and masked entity recovery steps, followed by masked multi-head attention and feed forward layers. Also includes fine-tuning tasks like entity linking, column type annotation, and cell filling.]

[Deng et al., VLDB’21]
**Type and position embeddings to mark different parts of a table and their relative positions**

- Reuse pre-trained embeddings when possible (e.g., initialized with TinyBERT [Jiao et al., 2020])
- Each entity has a unique entity embedding and one mention embedding (obtained from its surface form in the table)
Pre-training objectives:

- **Masked Language Model (MLM):** predict masked tokens in table metadata (i.e., table caption + headers)
- **Masked Entity Recovery (MEM):** predict the entity in a masked table cell
TURL: Table Understanding through Representation Learning

[Deng et al., VLDB’21]

• Pre-training data
  • Entity focused relational tables from Wikipedia that contains factual knowledge
  • 570171 / 5036 / 4964 tables for pre-training / validation / testing

<table>
<thead>
<tr>
<th>split</th>
<th>min</th>
<th>mean</th>
<th>median</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>1</td>
<td>13</td>
<td>8</td>
<td>4670</td>
</tr>
<tr>
<td>dev</td>
<td>5</td>
<td>20</td>
<td>12</td>
<td>667</td>
</tr>
<tr>
<td>test</td>
<td>5</td>
<td>21</td>
<td>12</td>
<td>3143</td>
</tr>
</tbody>
</table>

| # row  | train | 1   | 2 | 2  | 20  |
|        | dev   | 3   | 4 | 3  | 15  |
|        | test  | 3   | 4 | 3  | 15  |

| # ent. columns | train | 3   | 19 | 9  | 3911 |
|                | dev   | 8   | 57 | 34 | 2132 |
|                | test  | 8   | 60 | 34 | 9215 |

Dataset statistics (per table) in pre-training
TURL: Table Understanding through Representation Learning

• Pre-training data
  • Entity focused relational tables from Wikipedia that contains factual knowledge
  • 570171 / 5036 / 4964 tables for pre-training / validation / testing

[Deneg et al., VLDB’21]

With TinyBERT model architecture, smaller model, smaller pre-training datasets, greener data science
TURL: Table Understanding through Representation Learning

- Fine-tuning strategy for 6 different tasks

Pre-trained TURL can be applied to all 6 tasks with minimal modification.
• Experiments
  • 6 tasks on publicly available datasets as well as new datasets created in the paper
  • For example, *row population*:

  **Definition**: Given a partial table, and an optional set of seed subject entities, row population aims to retrieve more entities to fill the subject column.

  ![Input Table](image)

  ![Output Table](image)

  See more: [https://iai-group.github.io/webtables-tutorial/slides/part-5.pdf](https://iai-group.github.io/webtables-tutorial/slides/part-5.pdf)
TURL: Table Understanding through Representation Learning


• 6 tasks on publicly available datasets as well as new datasets created in the paper
• For example, **row population**:

<table>
<thead>
<tr>
<th># seed</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
<td>MAP</td>
<td>Recall</td>
</tr>
<tr>
<td>EntiTables [Zhang et al., 2017]</td>
<td>17.90</td>
<td>63.30</td>
</tr>
<tr>
<td>Table2Vec [Deng et al., 2019]</td>
<td>-</td>
<td>63.30</td>
</tr>
<tr>
<td>TURL + fine-tuning</td>
<td><strong>40.92</strong></td>
<td>63.30</td>
</tr>
</tbody>
</table>

All methods share the same candidate generation module, hence the same recall.
TURL: Table Understanding through Representation Learning

• Experiments
  • 6 tasks on publicly available datasets as well as new datasets created in the paper
  • For example, entity linking:

**Definition**: Given a table $T$ and a knowledge base KB, entity linking aims to link each potential mention in cells of $T$ to its referent entity $e \in KB$.

https://en.wikipedia.org/wiki/15th_National_Film_Awards
From Tables to Knowledge (KDD’21): Jay Pujara, Pedro Szekely, Huan Sun, Muhao Chen

TURL: Table Understanding through Representation Learning

• Experiments
  • 6 tasks on publicly available datasets as well as new datasets created in the paper
  • For example, entity linking:

Takeaways:
• Outperforms all baselines [Ritze et al., 2015, Efthymiou et al., 2017] on the challenging WikiGS [Efthymiou et al., 2017] and their new Test Set.

<table>
<thead>
<tr>
<th>Method</th>
<th>WikiGS</th>
<th>Our Test Set</th>
<th>T2D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>T2K [35]</td>
<td>34</td>
<td>70</td>
<td>22</td>
</tr>
<tr>
<td>Hybrid II [16]</td>
<td>64</td>
<td>69</td>
<td>60</td>
</tr>
<tr>
<td>Wikidata Lookup</td>
<td>57</td>
<td>67</td>
<td>49</td>
</tr>
<tr>
<td>TURL + fine-tuning w/o entity desc.</td>
<td>67</td>
<td>79</td>
<td>58</td>
</tr>
<tr>
<td>TURL + fine-tuning w/o entity type + reweighting</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>WikiLookup (Oracle)</td>
<td>74</td>
<td>88</td>
<td>64</td>
</tr>
</tbody>
</table>

[Deng et al., VLDB’21]
TURL: Table Understanding through Representation Learning

• Experiments
  • 6 tasks on publicly available datasets as well as new datasets created in the paper
  • For example, entity linking:

Takeaways:
• Outperforms all baselines [Ritze et al., 2015, Efthymiou et al., 2017] on the challenging WikiGS [Efthymiou et al., 2017] and their new Test Set.
• Generalizes well to general Web Tables (T2D [Lehmberg et al., 2016]). Improves Wikidata Lookup and obtain similar performance as state-of-the-art models.

<table>
<thead>
<tr>
<th>Method</th>
<th>WikiGS</th>
<th>Our Test Set</th>
<th>T2D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>T2K [35]</td>
<td>34</td>
<td>70</td>
<td>22</td>
</tr>
<tr>
<td>Hybrid II [16]</td>
<td>64</td>
<td>69</td>
<td>60</td>
</tr>
<tr>
<td>Wikidata Lookup</td>
<td>57</td>
<td>67</td>
<td>49</td>
</tr>
<tr>
<td>TURL + fine-tuning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o entity desc.</td>
<td>67</td>
<td>79</td>
<td>58</td>
</tr>
<tr>
<td>w/o entity type</td>
<td>60</td>
<td>70</td>
<td>52</td>
</tr>
<tr>
<td>+ reweighting</td>
<td>66</td>
<td>78</td>
<td>57</td>
</tr>
<tr>
<td>WikiLookup (Oracle)</td>
<td>74</td>
<td>88</td>
<td>64</td>
</tr>
</tbody>
</table>

[Deng et al., VLDB’21]
From Tables to Knowledge (KDD’21): Jay Pujara, Pedro Szekely, Huan Sun, Muhao Chen

TURL: Table Understanding through Representation Learning

• Summary:

(I) Learning deep *contextualized* representations of table elements
  • With a focus on relational tables (to be extended)
  • Modeling factual knowledge about named entities (Masked Entity Recovery)
  • Masked self-attention mechanism to model table structure

(II) Pre-training / fine-tuning paradigm works well for table-based tasks and greatly reduce feature engineering effort

(III) A benchmark (new datasets for 6 tasks) that is important for model development and evaluation

[Deng et al., VLDB’21]
Outline: Neural Representation Learning on Tables

• Background

• Representative Methods
  • Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR’19 (short); University of Stavanger]
  • TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB’21; OSU & Google]
  • TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL’21; Sony Co. & Adobe Research & UMass Amherst]
  • TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD’21; MSR & CMU & PKU]

• Summary

• Resources
TABBIE: Pretrained Representations of Tabular Data

- Idea:
  - Using the pretraining objective Corrupt Cell Detection [Clark et al., 2020]
TABBIE: Pretrained Representations of Tabular Data

• Model:
  • At one layer:
    • Two transformers acting row-wise and column-wise respectively

[1ida et al., NAACL’21]
TABBIE: Pretrained Representations of Tabular Data

- **Model:**
  - At one layer:
    - Two transformers acting row-wise and column-wise respectively
    - Average row and column result to create cell representation and pass to next layer

---

**Step 1:** compute *column* and *row* embeddings using two separate Transformers

**Step 2:** compute contextualized cell embeddings by averaging row/col embeddings

**Step 3:** feed these contextualized cell embeddings as input to the next layer

x12 layers

*Iida et al., NAACL’21*
### TABBIE: Pretrained Representations of Tabular Data

#### Tasks
- **Row population**
- **Column population**
- **Cell type annotation**

[![体育馆](image)](image)

<table>
<thead>
<tr>
<th>N</th>
<th>Method</th>
<th>MAP</th>
<th>MRR</th>
<th>Ndcg-10</th>
<th>Ndcg-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Entitables</td>
<td>36.8</td>
<td>45.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>TaBERT</td>
<td>43.2</td>
<td>55.7</td>
<td>45.6</td>
<td>47.7</td>
</tr>
<tr>
<td></td>
<td>TABBIE (FREQ)</td>
<td>42.8</td>
<td>54.2</td>
<td>44.8</td>
<td>46.9</td>
</tr>
<tr>
<td></td>
<td>TABBIE (MIX)</td>
<td>42.6</td>
<td>54.7</td>
<td>45.1</td>
<td>46.8</td>
</tr>
<tr>
<td>2</td>
<td>Entitables</td>
<td>37.2</td>
<td>45.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>TaBERT</td>
<td>43.8</td>
<td>56.0</td>
<td>46.4</td>
<td>48.8</td>
</tr>
<tr>
<td></td>
<td>TABBIE (FREQ)</td>
<td>44.4</td>
<td>57.2</td>
<td>47.1</td>
<td>49.5</td>
</tr>
<tr>
<td></td>
<td>TABBIE (MIX)</td>
<td>43.7</td>
<td>55.7</td>
<td>46.2</td>
<td>48.6</td>
</tr>
<tr>
<td>3</td>
<td>Entitables</td>
<td>37.1</td>
<td>44.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>TaBERT</td>
<td>42.9</td>
<td>55.1</td>
<td>45.6</td>
<td>48.5</td>
</tr>
<tr>
<td></td>
<td>TABBIE (FREQ)</td>
<td>43.4</td>
<td>56.5</td>
<td>46.6</td>
<td>49.0</td>
</tr>
<tr>
<td></td>
<td>TABBIE (MIX)</td>
<td>42.9</td>
<td>55.5</td>
<td>45.9</td>
<td>48.3</td>
</tr>
</tbody>
</table>

**Row population. N: number of seed rows**

[Barai et al., NAACL’21]
Outline: Neural Representation Learning on Tables

• Background

• Representative Methods
  • Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR’19 (short); University of Stavanger]
  • TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB’21; OSU & Google]
  • TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL’21; Sony Co. & Adobe Research & UMass Amherst]
  • TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD’21; MSR & CMU & PKU]

• Summary

• Resources
**TUTA: Tree-based Transformers for Generally Structured Table Pre-training**

[Jay Pujara, Pedro Szekely, Huan Sun, Muhao Chen]

**Goal**
- To encode generally structured tables

**Tables of varying structure**

---

(a) A vertical relational web table in Wikipedia

<table>
<thead>
<tr>
<th>Island</th>
<th>Nickname</th>
<th>Area</th>
<th>Population as of 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hawai’i</td>
<td>The Big Island</td>
<td>4,028.0 sq mi (10,432.5 km²)</td>
<td>185,079</td>
</tr>
<tr>
<td>Maui</td>
<td>The Valley Isle</td>
<td>727.2 sq mi (1,883.4 km²)</td>
<td>144,444</td>
</tr>
<tr>
<td>O’ahu</td>
<td>The Gathering Place</td>
<td>596.7 sq mi (1,544.5 km²)</td>
<td>963,207</td>
</tr>
</tbody>
</table>

(b) A horizontal entity web table

<table>
<thead>
<tr>
<th>Method</th>
<th>WikiGS</th>
<th>Our Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2K [35]</td>
<td>34</td>
<td>70</td>
</tr>
<tr>
<td>Hybrid II [16]</td>
<td>64</td>
<td>69</td>
</tr>
<tr>
<td>Wikidata Lookup</td>
<td>57</td>
<td>67</td>
</tr>
<tr>
<td>TURL + fine-tuning</td>
<td>67</td>
<td>79</td>
</tr>
<tr>
<td>w/o entity desc.</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>w/o entity type + reweighting</td>
<td>66</td>
<td>78</td>
</tr>
<tr>
<td>WikiLookup (Oracle)</td>
<td>74</td>
<td>88</td>
</tr>
</tbody>
</table>

(c) A matrix PDF table in arXiv

<table>
<thead>
<tr>
<th>Cancer statistics in 2018</th>
<th>Incidence</th>
<th>Mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>Skin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melanoma of skin</td>
<td>150,698</td>
<td>137,025</td>
</tr>
<tr>
<td>Non-melanoma skin cancer</td>
<td>637,733</td>
<td>404,323</td>
</tr>
<tr>
<td>Urinary tract</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kidney and renal pelvis</td>
<td>254,507</td>
<td>148,755</td>
</tr>
<tr>
<td>Bladder</td>
<td>424,082</td>
<td>125,311</td>
</tr>
<tr>
<td>Respiratory system</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Larynx</td>
<td>154,977</td>
<td>22,445</td>
</tr>
<tr>
<td>Trachea, bronchus and lung</td>
<td>1,368,524</td>
<td>725,352</td>
</tr>
<tr>
<td>Mesothelioma</td>
<td>21,662</td>
<td>8,781</td>
</tr>
</tbody>
</table>

(d) A matrix spreadsheet table
TUTA: Tree-based Transformers for Generally Structured Table Pre-training

- Bi-dimensional coordinate tree of a table

(Wang et al., SIGKDD’21)
TUTA: Tree-based Transformers for Generally Structured Table Pre-training

[Wang et al., SIGKDD’21]

Model highlights:

- **Model Input**
  - Numerical features
  - Top/left tree position
  - Formatting features
TUTA: Tree-based Transformers for Generally Structured Table Pre-training

[Wang et al., SIGKDD’21]

Model highlights:

- **Model Input**
  - Numerical features
  - Top/left tree position
  - Formatting features

- **Tree-based attention**
  - Only attend to structural neighborhood, defined base on tree distance
TUTA: Tree-based Transformers for Generally Structured Table Pre-training

[Wang et al., SIGKDD’21]

Model highlights:

• Model Input
  • Numerical features
  • Top/left tree position
  • Formatting features

• Tree-based attention
  • Only attend to structural neighborhood, defined base on tree distance
TUTA: Tree-based Transformers for Generally Structured Table Pre-training

• Three Pre-training Objectives

<table>
<thead>
<tr>
<th>Masked language modeling: (I)</th>
<th>Cell-level cloze: (II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict randomly masked tokens from the whole vocabulary</td>
<td>Fill the randomly masked cell locations from cell strings:</td>
</tr>
<tr>
<td>A. Mortality</td>
<td>A. Mortality</td>
</tr>
<tr>
<td>B. Urinary tract</td>
<td>B. Urinary tract</td>
</tr>
<tr>
<td>C. Mesothelioma</td>
<td>C. Mesothelioma</td>
</tr>
<tr>
<td>D. Kidney and renal pelvis</td>
<td>D. Kidney and renal pelvis</td>
</tr>
<tr>
<td>E. Colorectum and anus</td>
<td>E. Colorectum and anus</td>
</tr>
<tr>
<td>F. 725,352</td>
<td>F. 725,352</td>
</tr>
</tbody>
</table>

Table context retrieval: (III)

Retrieve table titles and descriptions from randomly selected text snippets:

- Family Households by Size, Type, and Age of Householder.
- It includes major cancer types including skin, respiratory, etc...
- Full-time Law Enforcement Officers.

---

From Tables to Knowledge (KDD’21): Jay Pujara, Pedro Szekely, Huan Sun, Muhao Chen
Pre-training Objectives

- Masked Language Model (MLM)
- Cell-level cloze: Map cell location to cell string
- Table context retrieval: Match cell representation (masked) to positive and negative candidate segments.

Tasks (Table Structure Understanding)

- Cell type classification (metadata, notes, data, top attribute, etc.)
- Table type classification (relational, list, entity, etc.)
Outline: Neural Representation Learning on Tables

• Background

• Representative Methods
  • Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval
    [Deng et al., SIGIR’19 (short); University of Stavanger]
  • TURL: Table Understanding through Representation Learning
    [Deng et al., VLDB’21; OSU & Google]
  • TABBIE: Pretrained Representations of Tabular Data
    [Iida et al., NAACL’21; Sony Co. & Adobe Research & UMass Amherst]
  • TUTA: Tree-based Transformers for Generally Structured Table Pre-training
    [Wang et al., SIGKDD’21; MSR & CMU & PKU]

• Summary

• Resources
## Summary

<table>
<thead>
<tr>
<th>Designs</th>
<th>Table2Vec</th>
<th>TURL</th>
<th>TUTA</th>
<th>TABBIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backbone Alg./Model</td>
<td>Word2Vec</td>
<td>Transformer with</td>
<td>Transformer with tree-based</td>
<td>Transformer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>masked self-attention</td>
<td>attention</td>
<td>(row-wise, column-wise)</td>
</tr>
<tr>
<td>Pretraining Objectives</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Masked Language Model (as</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>in BERT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell-level Cloze</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Table Context Retrieval</td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Masked Entity Recovery</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrupt Cell Detection</td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretraining Corpus</td>
<td>WikiTable: 1.6M tables</td>
<td>WikiTable: <strong>570K</strong> tables</td>
<td>WikiTable + WDC WebTable + web-crawled spreadsheet: <strong>57.9M</strong> tables</td>
<td>WikiTable + Common Crawl: 26.6M tables</td>
</tr>
<tr>
<td>Designs</td>
<td>Table2Vec</td>
<td>TURL</td>
<td>TUTA</td>
<td>TABBIE</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------</td>
<td>------------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>Downstream Tasks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entity Linking</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column Type Annotation</td>
<td>Yes</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Relation Extraction</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Row Population</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Cell Filling</td>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schema Augmentation</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column Population</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Table Type Classification</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Type Classification</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrupt Cell Detection</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fine-tuning</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
## Summary

- **Methods for Joint Representation Learning of Text and Tables**

<table>
<thead>
<tr>
<th>Method</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAPAS: Weakly Supervised Table Parsing via Pre-training</td>
<td>[Herzig et al., 2020]</td>
</tr>
<tr>
<td>StruG: Structure-Grounded Pretraining for Text-to-SQL</td>
<td>[Deng et al., 2021]</td>
</tr>
<tr>
<td>TABERT: Pretraining for Joint Understanding of Textual and Tabular Data</td>
<td>[Yin et al., 2020]</td>
</tr>
<tr>
<td>GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing</td>
<td>[Yu et al., 2021]</td>
</tr>
</tbody>
</table>
Summary

• Future directions
  • Systematic and fair comparison among different methods
    • For different types of tables
    • For different tasks
    • Easy-to-use benchmark
    • Considering pre-training data size, model size, training/inference time, etc.
    • Requires open-source code and datasets

• Better model architecture & pre-training objectives overall

• Joint representation of text, tables, and knowledge bases
Resources

• **Table2Vec**: [https://github.com/iai-group/sigir2019-table2vec](https://github.com/iai-group/sigir2019-table2vec)

• **TURL**: [https://github.com/sunlab-osu/TURL](https://github.com/sunlab-osu/TURL)
  (including pre-processed pre-training data, pre-trained model, new datasets for downstream tasks)

• **TABBIE**: [https://github.com/SFIG611/tabbie](https://github.com/SFIG611/tabbie)

• **TUTA**: [https://github.com/microsoft/TUTA_table_understanding/](https://github.com/microsoft/TUTA_table_understanding/) (to be released after internal review)
References

- Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. Structure-aware pre-training for table understanding with tree-based transformers. In SIGKDD’21.
Thank you!
More Experiments from TURL

**Column Type and Relation Extraction**
- Classification based on column representation.
- Takeaways:
  - Outperforms Sherlock [Hulsebos et al. 2019] and BERT-based RE model.
  - Handle fine-grained types well (e.g., actor, citytown).

### Column Type Annotation (Overall)

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sherlock (only entity mention)</td>
<td>78.47</td>
<td>88.40</td>
<td>70.55</td>
</tr>
<tr>
<td>TURL + fine-tuning (only entity mention)</td>
<td>88.86</td>
<td>90.54</td>
<td>87.23</td>
</tr>
<tr>
<td>TURL + fine-tuning w/o table metadata</td>
<td>93.77</td>
<td>94.80</td>
<td>92.76</td>
</tr>
<tr>
<td>TURL + fine-tuning w/o learned embedding</td>
<td>92.69</td>
<td>92.75</td>
<td>92.63</td>
</tr>
<tr>
<td>TURL + fine-tuning only table metadata</td>
<td>90.24</td>
<td>89.91</td>
<td>90.58</td>
</tr>
<tr>
<td>TURL + fine-tuning only learned embedding</td>
<td>93.33</td>
<td>94.72</td>
<td>91.97</td>
</tr>
</tbody>
</table>

### Relation Extraction

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-based</td>
<td>90.94</td>
<td>91.18</td>
<td>90.69</td>
</tr>
<tr>
<td>TURL + fine-tuning (only table metadata)</td>
<td>92.13</td>
<td>91.17</td>
<td>93.12</td>
</tr>
<tr>
<td>TURL + fine-tuning w/o table metadata</td>
<td>94.91</td>
<td>94.57</td>
<td>95.25</td>
</tr>
<tr>
<td>TURL + fine-tuning w/o learned embedding</td>
<td>93.85</td>
<td>93.78</td>
<td>93.91</td>
</tr>
</tbody>
</table>

### Column Type Annotation (Detailed)

<table>
<thead>
<tr>
<th>Method</th>
<th>person</th>
<th>pro_athlete</th>
<th>actor</th>
<th>location</th>
<th>citytown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sherlock</td>
<td>96.85</td>
<td>74.39</td>
<td>29.07</td>
<td>91.22</td>
<td>55.72</td>
</tr>
<tr>
<td>TURL + fine-tuning only entity mention</td>
<td>99.71</td>
<td>91.14</td>
<td>74.85</td>
<td>99.32</td>
<td>79.72</td>
</tr>
<tr>
<td>TURL + fine-tuning w/o table metadata</td>
<td>98.44</td>
<td>87.11</td>
<td>58.86</td>
<td>96.59</td>
<td>60.13</td>
</tr>
<tr>
<td>TURL + fine-tuning w/o learned embedding</td>
<td>99.38</td>
<td>90.38</td>
<td>74.46</td>
<td>99.01</td>
<td>77.37</td>
</tr>
<tr>
<td>TURL + fine-tuning only table metadata</td>
<td>98.26</td>
<td>88.80</td>
<td>71.39</td>
<td>98.91</td>
<td>75.55</td>
</tr>
<tr>
<td>TURL + fine-tuning only learned embedding</td>
<td>98.72</td>
<td>91.06</td>
<td>73.62</td>
<td>97.78</td>
<td>75.16</td>
</tr>
</tbody>
</table>
More Experiments from TURL

- **Ablation study**
  - Masked self-attention (visibility matrix) is important for modeling table structure.
  - MER mask ratio can affect the model performance.

[Graph showing effect of visibility matrix and different MER mask ratios on top-1 object entity prediction accuracy on validation set.]

[Deng et al., VLDB’21]