Team Project HWS 2022

Table Annotation using Deep Learning
− **Prof. Dr. Christian Bizer**
− Professor for Information Systems V
− Research Interests:
  • Web Data Integration
  • Data and Web Mining
  • Deployment of Data Web Technologies
− Room: B6 - B1.15
− eMail: christian.bizer@uni-mannheim.de
− Consultation: Wednesday, 13:30-14:30
Hallo

- **M. Sc. Wi-Inf. Keti Korini**
- Graduate Research Associate
- Research Interests:
  - Table Annotation using Deep Learning
  - Schema Matching
- Room: B6, 26, C 1.03
- eMail: kkorini@uni-mannheim.de
Agenda of Today’s Kickoff Meeting

1. A round of introductions: You and Your Experience
2. Introduction and Project Goals
3. The WDC Schema.org Table Annotation Benchmark
4. Organization
5. Specific Subtasks
6. Schedule
7. Formal Requirements
You and Your Experience

- A Short Round of Introductions
  - What are you studying? Which semester?
  - Which DWS courses did you already attend?
  - What are your programming and data wrangling skills?
  - Did you already work on any data integration or cleansing projects?

- Participants
  1. Poudel, Subash
  2. Der, Reng
  3. Chen, Chun-Yi
  4. Hsieh, I-Chen
  5. Abobaker, Munir
2. Introduction: The Web as Source of Structured Data

There are hundreds of millions of high-quality tables available on the Web and in Wikipedia.
The Web as Source of Structured Data

In addition, there are millions of tables available via public data portals.
Challenge for Data Search

For providing advanced dataset search, one needs to understand the schemata of the tables.

Give me datasets describing movies that provide directors and release years?

Give me all datasets containing population numbers for German cities?
Challenge for Data Integration

For using data from these tables to complete databases or knowledge graphs, one needs to understand the schema of the tables.
Table Annotation

Goal: Annotate tables in a **large table corpus** with concepts from a **knowledge graph or shared vocabulary**

- Enabling step for data search and data integration
- Challenges: Heterogeneous or unknown headers, heterogeneous content

<table>
<thead>
<tr>
<th>film</th>
<th>director</th>
<th>producer</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy Feet</td>
<td>George Miller, Warren Coleman, Judy Morris</td>
<td>Bill Miller, George Miller, Doug Mitchell</td>
<td>USA</td>
</tr>
<tr>
<td>Cars</td>
<td>John Lasseter, Joe Ranft</td>
<td>Darla K. Anderson</td>
<td>UK</td>
</tr>
<tr>
<td>Flushed Away</td>
<td>David Bowers, Sam Fell</td>
<td>Dick Clement, Ian La Frenais, Simon Nye</td>
<td>France</td>
</tr>
</tbody>
</table>
Table Annotation Tasks

1. cell entity annotation (CEA)
2. column type annotation (CTA)
3. columns property annotation (CPA)
4. row annotation
5. table type detection

Papers with Code: Table Annotation: https://paperswithcode.com/task/table-annotation
Task 1: Column Type Annotation

Column Type Annotation (CTA): Annotation of table columns with the type of the entities contained in the column.

Entity types capture more domain semantics than data types:

- **Data types:** string, integer, Boolean …
- **Entity types:** country, product, person, duration, distance, weight, …

→ Usually approached as a multi-class classification problem

- **Input:** column + table context
- **Output:** Column type label from set of possible labels
Task 2: Columns Property Annotation

**Columns Property Annotation (CPA):** Annotation of the relationship between the subject column and a second column.

Subject/label column contains the name of the entity described in a row:
- often the left-most column in the table,
- the other columns describe attributes of the named entity

→ Treated as **multi-class classification** problem
- **Input:** column pair + table context
- **Output:** columns property label
Transformer architectures like BERT had large impact on NLP

- stacked encoder layers with self-attention mechanism
- contextual representation of tokens within a sequence of tokens
- [CLS] token that summarizes the sequence and can be used for classification

Pre-training / Fine-tuning paradigm

- pre-training: self-supervised masked language modeling on large text-corpora
- fine-tuning: further training on task-specific data

→ shown to work extremely well for a variety of problems
→ growing body of work regarding data integration using Transformers (see references)
Transformers for Entity Matching: DITTO (2021)

- applies BERT, DistilBERT, RoBERTa for entity matching
- adds methods for entity summarization, training data augmentation
- Entity serialization for BERT
  - Pair of entity descriptions are turned into single sequence
  - [CLS] Entity Description 1 [SEP] Entity Description 2 [SEP]
  - Entity Description = [COL] attr_1 [VAL] val_1 . . . [COL] attr_k [VAL] val_k
- [CLS] token summarizes the pair of entities
- linear layer on top of [CLS] token for matching decision
- uses augmentation to increase amount of training data

### DITTO: Evaluation

<table>
<thead>
<tr>
<th>Type</th>
<th>Dataset</th>
<th>DITTO F1</th>
<th>DeepMatcher F1</th>
<th>Magellan F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>iTunes-Amazon</td>
<td>97.0</td>
<td>88.5 +8.5</td>
<td>91.2 +5.8</td>
</tr>
<tr>
<td></td>
<td>DBLP-ACM</td>
<td>99.0</td>
<td>98.4 +0.6</td>
<td>98.4 +0.6</td>
</tr>
<tr>
<td></td>
<td>DBLP-Scholar</td>
<td>95.6</td>
<td>92.3 +3.3</td>
<td>94.7 +0.9</td>
</tr>
<tr>
<td></td>
<td>Walmart-Amazon</td>
<td>86.8</td>
<td>66.9 +19.9</td>
<td>71.9 +14.9</td>
</tr>
<tr>
<td></td>
<td>Abt-Buy</td>
<td>89.3</td>
<td>62.8 +26.5</td>
<td>43.6 +45.7</td>
</tr>
<tr>
<td></td>
<td>Amazon-Google</td>
<td>75.6</td>
<td>69.3 +6.3</td>
<td>49.1 +26.5</td>
</tr>
<tr>
<td>Textual</td>
<td>WDC Computer - Large</td>
<td>91.7</td>
<td>89.5 +3.2</td>
<td>64.5 +27.2</td>
</tr>
<tr>
<td></td>
<td>WDC Computer - Small</td>
<td>80.8</td>
<td>70.5 +10.3</td>
<td>57.6 +23.2</td>
</tr>
</tbody>
</table>

- constant improvement for structured data
- large performance gain for textual data
Potential Reasons for the Performance Gain

- Serialization allows to pay attention to all attributes
  - no strict separation between attributes
- WordPiece tokenizer breaks unknown terms into pieces
  - no problems with out of vocabulary terms
- Transfer learning from pre-training texts
  - different surface forms are already close in embedding space
- Contextualization of the embeddings
  - potentially more suited for capturing differing semantics
Transformers for Table Annotation

  - uses TinyBERT as base architecture, pre-trained on text corpora
  - Additional pre-training on ~600k Wikipedia tables using **Masked Language Modeling** and a **Masked Entity Recovery** objective
  - Fine-tuned and evaluated on entity linking (CEA), CTA, CPA, row population and cell filling

1. Pre-training on text

2. Pre-training on structured data

3. Fine-tuning for specific tasks

Transformers for Table Annotation

- **Visibility Matrix** = attention mask during self-attention calculation
  - Determines which other tokens from the other table cells are visible for each token

Transformers for Table Annotation

  - fine-tunes BERT (no further pre-training) for CTA and CPA and uses multi-task learning
  - evaluation of single-column and multi-column models

Evaluation Results of Table Annotation Systems on WikiTables

- Column Type Annotation (~255 types)

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>TURL (TinyBERT)</td>
<td>88.86</td>
<td>90.54</td>
<td>87.23</td>
</tr>
<tr>
<td>DoDuo (BERT)</td>
<td>92.45</td>
<td>92.45</td>
<td>92.21</td>
</tr>
</tbody>
</table>

- Column Property Annotation (~121 relations)

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>TURL (TinyBERT)</td>
<td>90.94</td>
<td>91.18</td>
<td>90.69</td>
</tr>
<tr>
<td>DoDuo (BERT)</td>
<td>91.72</td>
<td>91.97</td>
<td>91.47</td>
</tr>
</tbody>
</table>

Further Table Embedding Approaches

  - general pre-training on 1.8M Wikipedia Tables and 24.8M WebTables
  - using a cell corruption objective: Replace some cell contents with frequency-based cell sampling across all tables and then try to predict if a cell was changed or not
  - evaluated on Column Type Annotation, Row Population, Column Population
  - code available

  - pre-trained on relational and spreadsheet tables from Wikipedia and the WDC WebTable Corpus
  - using a Masked Language Model, multi-choice Cloze at the cell level and context retrieval at table level
  - evaluated on Cell Type and Table Type Classification
  - code not available

Evaluation Campain for Table to Knowledge Graph Matching

- Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)
- Yearly challenge where table annotation methods are compared on their performance regarding the **CTA**, **CPA** and **CEA** tasks on the same benchmark datasets
- Table columns and cells are linked to a Knowledge Graph (such as DBpedia or WikiData) concept

Image Source: N. Abdelmageed, et al.: BiodivTab: A Table Annotation Benchmark based on Biodiversity Research Data

Benchmarks using Annotated Tables as Training Data

- **WikiTables**
  - 580,171 tables extracted from Wikipedia
  - 406,706 of these tables are labeled for the Column Type Annotation (CTA) task (255 labels)
  - 55,970 tables labeled for the Columns Property Annotation (CPA) task (121 labels)
  - labels are taken from Freebase knowledge graph
  - Task: Predict the CTA and CPA labels for tables where the true labels are hidden

- **Schema.org Table Annotation Benchmark (SOTAB)**
  - tables containing schema.org data originating from 74,215 different websites
  - 59,548 tables annotated for CTA (91 labels)
  - 48,379 tables annotated for CPA (176 labels)
  - labels are mostly taken from the schema.org vocabulary
  - Task: Predict the CTA and CPA labels for tables where the true labels are hidden
Project Goals

1. Try to improve the state-of-the-art performance for the CTA and CPA tasks on the SOTAB benchmark, by experimenting with
   1. different **table serialization** techniques
   2. **table augmentation** techniques
   3. different methods for embedding table columns plus context

2. Compare performance of your methods to
   - existing table annotation methods (e.g. DoDuo, TURL, …) using different evaluation datasets

3. Try to explain why models perform better than others
   - evaluate performance on different challenge columns
   - conduct an error analysis
Learning Targets

Improve your technical skills
- Work as a **Data Scientist**: clean, profile, classify data
- Understand the nature of **Web Data**
- Improve your technical expertise concerning **Deep Learning**
- Improve your programming skills

Improve your soft skills
- Work as part of a bigger team on a more complex project
- Organize yourself and assign tasks based on your skills
- Communicate and coordinate your work
The WDC Schema.Org Table Annotation Benchmark

1. Semantic Annotations in HTML Pages
2. Web Data Commons – Schema.org Table Corpus
3. Schema.org Table Annotation Benchmark (SOTAB)
Semantic Annotation of HTML Pages: Schema.org

- ask site owners since 2011 to annotate data for enriching search results
- 675 Types: Event, Place, Local Business, Product, Review, Person
- Encoding: Microdata, RDFa, JSON-LD
Example: Microdata Annotations in HTML

```html
<div itemtype="http://schema.org/Product">
  <span itemprop="name">Sony GTK-XB5L Audiosystem</span>
  <span itemprop="gtin13">04048945021687</span>
  <span itemprop="description">high-power home audio system with Bluetooth technology</span>
</div>

<div itemprop="aggregateRating" itemscope itemtype="http://schema.org/AggregateRating">
  <span itemprop="ratingValue">4</span> stars-based on
  <span itemprop="reviewCount">250</span> reviews.
</div>
```
# Frequently used Schema.org Classes (2020)

<table>
<thead>
<tr>
<th>Class</th>
<th># Websites (PLDs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>schema:WebPage</td>
<td>4,484,026</td>
</tr>
<tr>
<td></td>
<td>1,339,999</td>
</tr>
<tr>
<td>schema:Person</td>
<td>3,151,809</td>
</tr>
<tr>
<td></td>
<td>514,990</td>
</tr>
<tr>
<td>schema:BreadcrumbList</td>
<td>1,688,820</td>
</tr>
<tr>
<td></td>
<td>924,991</td>
</tr>
<tr>
<td>schema:Article</td>
<td>1,327,578</td>
</tr>
<tr>
<td></td>
<td>627,303</td>
</tr>
<tr>
<td>schema:Product</td>
<td>1,234,972</td>
</tr>
<tr>
<td></td>
<td>1,059,149</td>
</tr>
<tr>
<td>schema:Offer</td>
<td>1,182,855</td>
</tr>
<tr>
<td></td>
<td>946,725</td>
</tr>
<tr>
<td>schema:PostalAddress</td>
<td>863,243</td>
</tr>
<tr>
<td></td>
<td>585,417</td>
</tr>
<tr>
<td>schema:BlogPosting</td>
<td>529,020</td>
</tr>
<tr>
<td></td>
<td>552,338</td>
</tr>
<tr>
<td>schema:LocalBusiness</td>
<td>363,843</td>
</tr>
<tr>
<td></td>
<td>280,338</td>
</tr>
<tr>
<td>schema:AggregateRating</td>
<td>432,014</td>
</tr>
<tr>
<td></td>
<td>315,253</td>
</tr>
<tr>
<td>schema:Place</td>
<td>255,139</td>
</tr>
<tr>
<td></td>
<td>93,124</td>
</tr>
<tr>
<td>schema:Event</td>
<td>194,115</td>
</tr>
<tr>
<td></td>
<td>77,722</td>
</tr>
<tr>
<td>schema:Review</td>
<td>181,097</td>
</tr>
<tr>
<td></td>
<td>158,333</td>
</tr>
<tr>
<td>schema:JobPosting</td>
<td>28,759</td>
</tr>
<tr>
<td></td>
<td>8,520</td>
</tr>
</tbody>
</table>

http://webdatacommons.org/structureddata/2020-12/stats/schema_org_subsets.html
## Properties used to Describe Products (2020)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>% of PLDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>schema:Product/name</td>
<td>99 %</td>
</tr>
<tr>
<td>schema:Product/offers</td>
<td>94 %</td>
</tr>
<tr>
<td>schema:Offer/price</td>
<td>95 %</td>
</tr>
<tr>
<td>schema:Offer/priceCurrency</td>
<td>95 %</td>
</tr>
<tr>
<td>schema:Product(description)</td>
<td>84 %</td>
</tr>
<tr>
<td>schema:Offer/availability</td>
<td>72 %</td>
</tr>
<tr>
<td>schema:Product/sku</td>
<td>56 %</td>
</tr>
<tr>
<td>schema:Product/brand</td>
<td>30 %</td>
</tr>
<tr>
<td>schema:Product/image</td>
<td>26 %</td>
</tr>
<tr>
<td>schema:Product/aggregateRating</td>
<td>17 %</td>
</tr>
<tr>
<td>schema:Product/mpn</td>
<td>6.3 %</td>
</tr>
<tr>
<td>schema:Product/productID</td>
<td>4.7 %</td>
</tr>
</tbody>
</table>

The Galaxy S4 is among the earliest phones to feature a 1080p Full HD display. The various connectivity options on the Samsung include …

http://webdatacommons.org/structureddata/schemaorgtables/
The WDC Schema.Org Table Corpus

- Extract annotations for 43 schema.org entity classes
- Extract attribute values and group entities by website
- Initial cleaning steps are performed
- **Final result**: 4.2M tables - one table per domain, containing all annotated entities after cleaning.
- All tables share the same schema

Details and Download:
- [http://webdatacommons.org/structureddata/schemaorgtables/](http://webdatacommons.org/structureddata/schemaorgtables/)
The WDC Schema.org Table Annotation Benchmark

- covers the **Column Type Annotation** and **Column Property Annotation** tasks
- uses tables from the in Schema.org Table Corpus
  - tables with schema.org labels are given for training
  - tables with masked-out labels are used for testing
- Table columns selected based on three challenges:
  - **Missing values**: Columns where missing cells are present, between 10-70 % density
  - **Value Format Heterogeneity**: Columns that contain values with different (measurement) formats
  - **Corner Cases**: Columns that are hard to annotate, similar columns that have different labels, dissimilar columns that have the same label
  - **Random columns**: Randomly chosen columns for all chosen labels

→ **Result**: ~135,000 annotated columns and column relations from ~50,000 tables for Column Type Annotation and Column Property Annotation tasks separately
**Benchmark Statistics**

- **Label Space**: 91 semantic types for CTA and 176 relation labels for CPA.
  - Corresponding to 17 domains: Product, JobPosting, Hotel, Restaurant, Event, MusicAlbum, Person ...

<table>
<thead>
<tr>
<th>CPA Overall Top 10 Labels/Annotated Columns</th>
<th>CTA Overall Top 10 Labels/Annotated Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>description</td>
<td>Text</td>
</tr>
<tr>
<td>datePublished</td>
<td>DateTime</td>
</tr>
<tr>
<td>telephone</td>
<td>Duration</td>
</tr>
<tr>
<td>author</td>
<td>Mass</td>
</tr>
<tr>
<td>price</td>
<td>Date</td>
</tr>
<tr>
<td>startDate</td>
<td>currency</td>
</tr>
<tr>
<td>priceCurrency</td>
<td>Number</td>
</tr>
<tr>
<td>ratingValue</td>
<td>telephone</td>
</tr>
<tr>
<td>addressCountry</td>
<td>price</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>description</th>
<th>7824</th>
</tr>
</thead>
<tbody>
<tr>
<td>datePublished</td>
<td>4963</td>
</tr>
<tr>
<td>telephone</td>
<td>4731</td>
</tr>
<tr>
<td>author</td>
<td>4151</td>
</tr>
<tr>
<td>price</td>
<td>4132</td>
</tr>
<tr>
<td>startDate</td>
<td>3918</td>
</tr>
<tr>
<td>priceCurrency</td>
<td>3597</td>
</tr>
<tr>
<td>ratingValue</td>
<td>3546</td>
</tr>
<tr>
<td>addressCountry</td>
<td>3373</td>
</tr>
<tr>
<td>Text</td>
<td>7771</td>
</tr>
<tr>
<td>DateTime</td>
<td>6428</td>
</tr>
<tr>
<td>Duration</td>
<td>5997</td>
</tr>
<tr>
<td>Mass</td>
<td>5649</td>
</tr>
<tr>
<td>Date</td>
<td>5508</td>
</tr>
<tr>
<td>currency</td>
<td>5383</td>
</tr>
<tr>
<td>Number</td>
<td>5160</td>
</tr>
<tr>
<td>telephone</td>
<td>4647</td>
</tr>
<tr>
<td>price</td>
<td>4475</td>
</tr>
</tbody>
</table>
Benchmark Statistics

- Tables have a minimum of 10 rows and a minimum of 3 columns.
- Includes textual, numerical and datetime column values
- Fixed training, validation and testing splits
  - Training set is provided in two sizes: **Large** training set and **Small** training set

**Overall Top 5 Table Types in SOTAB**

<table>
<thead>
<tr>
<th>Type</th>
<th>Overall</th>
<th>Training Set</th>
<th>Small Training Set</th>
<th>Validation Set</th>
<th>Testing Set</th>
<th>Overall</th>
<th>Training Set</th>
<th>Small Training Set</th>
<th>Validation Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schema.org</td>
<td># tables</td>
<td># columns</td>
<td>median rows/cols</td>
<td># tables</td>
<td># columns</td>
<td># tables</td>
<td># columns</td>
<td># tables</td>
<td># columns</td>
<td># tables</td>
</tr>
<tr>
<td>Product</td>
<td>12,534</td>
<td>30,656</td>
<td>40/8</td>
<td>10,403</td>
<td>23,383</td>
<td>2,037</td>
<td>5,211</td>
<td>857</td>
<td>2,507</td>
<td>2,196</td>
</tr>
<tr>
<td>Event</td>
<td>6,256</td>
<td>17,972</td>
<td>23/7</td>
<td>4,721</td>
<td>14,150</td>
<td>1,386</td>
<td>4,631</td>
<td>593</td>
<td>1,669</td>
<td>1,353</td>
</tr>
<tr>
<td>LocalBusiness</td>
<td>5,292</td>
<td>14,941</td>
<td>37/9</td>
<td>3,744</td>
<td>10,657</td>
<td>1,226</td>
<td>3,932</td>
<td>602</td>
<td>2,161</td>
<td>2,123</td>
</tr>
<tr>
<td>Recipe</td>
<td>5,080</td>
<td>24,590</td>
<td>42/13</td>
<td>4,083</td>
<td>20,809</td>
<td>766</td>
<td>4,459</td>
<td>412</td>
<td>2,199</td>
<td>1,582</td>
</tr>
<tr>
<td>CreativeWork</td>
<td>4,674</td>
<td>9,240</td>
<td>26/4</td>
<td>3,393</td>
<td>6,521</td>
<td>1,142</td>
<td>2,299</td>
<td>665</td>
<td>1,355</td>
<td>964</td>
</tr>
<tr>
<td>Person</td>
<td>4,623</td>
<td>10,466</td>
<td>27/5</td>
<td>3,818</td>
<td>8,850</td>
<td>851</td>
<td>1,876</td>
<td>374</td>
<td>867</td>
<td>919</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Overall</th>
<th>Training Set</th>
<th>Small Training Set</th>
<th>Validation Set</th>
<th>Testing Set</th>
<th>Overall</th>
<th>Training Set</th>
<th>Small Training Set</th>
<th>Validation Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schema.org</td>
<td># tables</td>
<td># columns</td>
<td>median rows/cols</td>
<td># tables</td>
<td># columns</td>
<td># tables</td>
<td># columns</td>
<td># tables</td>
<td># columns</td>
<td># tables</td>
</tr>
<tr>
<td>Overall</td>
<td>59,548</td>
<td>162,351</td>
<td>33/7</td>
<td>46,790</td>
<td>130,471</td>
<td>11,517</td>
<td>33,004</td>
<td>5,732</td>
<td>16,840</td>
<td>7,026</td>
</tr>
<tr>
<td>Product</td>
<td>12,534</td>
<td>30,656</td>
<td>40/8</td>
<td>10,403</td>
<td>23,383</td>
<td>2,037</td>
<td>5,211</td>
<td>857</td>
<td>2,507</td>
<td>2,196</td>
</tr>
<tr>
<td>Event</td>
<td>6,256</td>
<td>17,972</td>
<td>23/7</td>
<td>4,721</td>
<td>14,150</td>
<td>1,386</td>
<td>4,631</td>
<td>593</td>
<td>1,669</td>
<td>1,353</td>
</tr>
<tr>
<td>LocalBusiness</td>
<td>5,292</td>
<td>14,941</td>
<td>37/9</td>
<td>3,744</td>
<td>10,657</td>
<td>1,226</td>
<td>3,932</td>
<td>602</td>
<td>2,161</td>
<td>2,123</td>
</tr>
<tr>
<td>Recipe</td>
<td>5,080</td>
<td>24,590</td>
<td>42/13</td>
<td>4,083</td>
<td>20,809</td>
<td>766</td>
<td>4,459</td>
<td>412</td>
<td>2,199</td>
<td>1,582</td>
</tr>
<tr>
<td>CreativeWork</td>
<td>4,674</td>
<td>9,240</td>
<td>26/4</td>
<td>3,393</td>
<td>6,521</td>
<td>1,142</td>
<td>2,299</td>
<td>665</td>
<td>1,355</td>
<td>964</td>
</tr>
<tr>
<td>Person</td>
<td>4,623</td>
<td>10,466</td>
<td>27/5</td>
<td>3,818</td>
<td>8,850</td>
<td>851</td>
<td>1,876</td>
<td>374</td>
<td>867</td>
<td>919</td>
</tr>
</tbody>
</table>

Universität Mannheim – Bizer/Korini: Team Project – HWS2022 – Slide 35
## Ground Truth Files

- **Tables and ground truth files provided**
  
  - **Tables** provided without column names
    
    | 0 | Designer Inspired Chain Necklace Glass Dome Pe... | 40th Chain Classic designer inspired famous fl... | USD | 69 | https://schema.org/InStock | Beauty In Stone Jewelry | 2020-10-07 |
    | 1 | Genuine Pearl With Modern Cube on Chain Necklace | Bridal Pearl Necklace Genuine freshwater pearl... | USD | 49 | https://schema.org/InStock | Beauty In Stone Jewelry | 2020-10-07 |
    | 2 | Beach Glass Necklace With Crystal Pendant | Dark gray beach glass, metal patina beads on n... | USD | 99 | https://schema.org/InStock | Beauty In Stone Jewelry | 2020-10-09 |
    | 3 | Brushed Gold Bracelet | 22k Gold Electroplate over brass beads. Stretc... | USD | 65 | https://schema.org/InStock | Beauty In Stone Jewelry | 2020-10-03 |

  - **Ground Truth files** for training, validation and 5 test sets: full test set, missing values test set, random test set, format heterogeneity test set and corner cases test set.

### CTA Ground Truth Example

<table>
<thead>
<tr>
<th>table_name</th>
<th>column_index</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Product_beautyinstonejewelry.com_September2020...</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>Product_beautyinstonejewelry.com_September2020...</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Product_beautyinstonejewelry.com_September2020...</td>
<td>5</td>
</tr>
</tbody>
</table>

### CPA Ground Truth Example

<table>
<thead>
<tr>
<th>table_name</th>
<th>main_column_index</th>
<th>column_index</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Product_corememoriesco.com_September2020_CPA.j...</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>Product_corememoriesco.com_September2020_CPA.j...</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Product_corememoriesco.com_September2020_CPA.j...</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>
Team Project Organization

**Duration:** 6 months (03.10.2022 – 03.04.2023)

**Participants:** 5 people

**Type of work:** Team and subgroup based

**Milestones:** 3 project phases

**ECTS Points:** 12

**Evaluation**

- Intermediate presentations
- Final report as HTML page
- Individual contribution to the deliverables is graded
Team Project Organization: Project Phases

First phase
- Column Type Annotation

Second phase
- Columns Property Annotation
- Refinement

Third phase
- Transfer and Report
Phase 1a: Column Type Annotation

Participants: all team members

Duration: 3.10 – 14.11

Input: SOTAB CTA dataset (*provided here*)

Sub-phases:

1. Topic understanding and ideas collection
   - Read papers, get ideas about *Table Augmentation* and *Table Serialization* and decide which methods to implement and try out
   - **Sub-phase duration:** 3.10 – 10.10
   - **Deliverable:** List of methods and description of each method
Table Serialization: Ideas where to start

- What is table serialization?
  - Language models take **token sequences** as input
  - **Table Tasks**: columns are converted into token sequences

- **Basic Serialization**: Serialize one column
  - Concatenate all column values
  - Add special [CLS] token at the beginning of sequence
  - Add special [SEP] token at the end of the sequence
    
    [CLS] Alofi Nice Dublin [SEP]

- **Table Serialization**: Serialize all table columns
  - Concatenate all column values and separate them with [CLS] tokens, add a [SEP] token at the end:
    [CLS] Lau’s Gateway Radisson ... [CLS] 209 Main Street ... [CLS] Alofi Nice .... [CLS] NZD ... [SEP]

- Think of **other serialization possibilities** that could be used to improve models:
  - Concatenate only neighboring columns or break down tables into smaller tables
  - Summarize long column values using ex. TF-IDF, if table is long sampling from values
Table Augmentation: Ideas where to start

- Try data augmentation techniques in the context of tables
- At different levels:
  - Cell level
  - Row/column level
  - Table level
- Examples:
  - **Cell level**: Delete cell values, Replace tokens ...
  - **Row/column level**: Shuffle column value or row value order
  - **Table level**: Mix rows and columns from different tables
- Can table augmentation improve performance results, especially using the small training set where less examples are available?
Phase 1a: Column Type Annotation

Participants: all team members

Duration: 3.10 – 14.11

Input: SOTAB CTA dataset (provided here)

Sub-phases:

2. Implement methods and run experiments
   - Baseline methods (Random Forest, SVM), BERT, RoBERTa, Doduo, Contrastive Learning
   - Sub-phase duration: 10.10 – 07.11
   - Deliverables: Code and Evaluation results

3. Analyze results and develop ideas for improvement for second iteration
   - Error analysis: Look at correctly/incorrectly classified examples for select models and calculate statistics for error classes
   - Sub-phase duration: 07.11 – 14.11
   - Deliverables: Error analysis report and plans for the next iteration
Phase 1a: Column Type Annotation - How to get started?

- Get the Tables ([here](#))
- Get acquainted with data, ground truth files, etc.
- Experiment with table augmentation and serialization using baseline methods such as BERT, or more advanced methods like DODUO ([https://github.com/megagonlabs/doduo](https://github.com/megagonlabs/doduo))
- Code to get started will be provided in Github in a shared repository
- We will provide access to GPU server in the DWS student server
- You can also use the BW Uni Cluster
  - [https://wiki.bwhpc.de/e/Category:BwUniCluster_2.0](https://wiki.bwhpc.de/e/Category:BwUniCluster_2.0)
Phase 1b: Columns Property Annotation

**Participants:** all team members

**Duration:** 14.11 – 19.12

**Input:** SOTAB CPA dataset (*provided here*)

**Sub-phases:**

1. Topic understanding and ideas collection
   - Read papers, get ideas about *Table Augmentation* and *Table Serialization* and decide what methods to use in the evaluation phase
   - **Sub-phase duration:** 14.11 – 21.11
   - **Deliverables:** List of ideas to try and description
Table Augmentation and Serialization: Ideas for CPA

- **Table Augmentation for CPA**: focus on augmentation of pairs of columns within the same table

- Some similar examples as in the CTA phase:
  - **Cell level**: Delete cell values, Replace tokens ...
  - **Row/column level**: Shuffle column value or row value order

- **Table Serialization for CPA**: Normally treats a column pair as a sequence
  - Concatenate all column values of each column
  - Add [CLS] token at the beginning and [SEP] token to separate the two columns and at the end of the sequence
  - Further idea: Add other columns as context
  - Summarize long textual columns (TF-IDF method)
Phase 1b: Columns Property Annotation

Participants: all team members

Duration: 14.11 – 19.12

Input: SOTAB CPA dataset (provided here)

Sub-phases:

2. Implement methods and run experiments
   • Baseline methods (Random Forest, SVM), BERT, RoBERTa, Doduo, Contrastive Learning
   • Sub-phase duration: 21.11 – 12.12
   • Deliverables: Code and Evaluation results

3. Analyze results and develop ideas for improvement for second iteration
   • Error analysis: Look at correctly/incorrectly classified examples for select models and calculate statistics for error classes
   • Sub-phase duration: 12.12 – 19.12
   • Deliverables: Error analysis report and plans for the next iteration
Phase 2: Refinement (Second Iteration)

Participants: two subgroups (one subgroup per CTA/CPA task)

Duration: 19.12 – 06.02

Sub-phases:

1. Improve code for CTA and CPA task and rerun experiments
   • Based on the plan for refinement made on the previous phase
   • **Sub-phase duration:** 19.12 – 16.01
   • **Deliverables:** Code and Evaluation Results

2. Error analysis
   • Look at correctly/incorrectly classified examples for select models
   • How did the models perform in challenge columns (Missing values...) and in random columns?
   • How did the models perform on different types of data (numerical, datetime, text)?
   • **Sub-phase duration:** 16.01 – 06.02
   • **Deliverables:** Error Analysis Report
Phase 3: Transfer and Report

Participants: all team members or two subgroups per task (CPA/CTA)

Duration: 06.02 – 03.04

Input: Results from Phase 2

Sub-phases:

1. Test good working approaches on other datasets
   • such as Wikitables by TURL and GitTables benchmark
   • include results from error analysis into SOTAB test and training
   • sub-phase duration: 06.02 – 06.03

2. Write report
   • as an HTML page
   • sample HTML code will be provided
   • sub-phase duration: 06.03 – 03.04
<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wednesday, 28.09.2022</td>
<td>Kickoff meeting (today)</td>
</tr>
<tr>
<td></td>
<td><strong>Phase 1a:</strong> CTA Topic Understanding and ideas collection</td>
</tr>
<tr>
<td>Monday, 10.10.2022</td>
<td>1\textsuperscript{st} Deliverable: List of methods and description of each method</td>
</tr>
<tr>
<td></td>
<td><strong>Phase 1a:</strong> CTA Experiments</td>
</tr>
<tr>
<td>Monday, 21.10.2022</td>
<td>Meet Keti and report current plan and results</td>
</tr>
<tr>
<td>Monday, 07.11.2022</td>
<td>2\textsuperscript{nd} Deliverable: Code and evaluation results</td>
</tr>
<tr>
<td></td>
<td><strong>Phase 1a:</strong> CTA improvements plan</td>
</tr>
<tr>
<td>Monday, 14.11.2022</td>
<td>3\textsuperscript{rd} Deliverable: Error analysis report and plans for the next iteration</td>
</tr>
<tr>
<td></td>
<td><strong>Phase 1b:</strong> CPA Topic Understanding and ideas collection</td>
</tr>
<tr>
<td>Monday, 21.11.2022</td>
<td>4\textsuperscript{th} Deliverable: List of methods and description of each method</td>
</tr>
<tr>
<td></td>
<td><strong>Phase 1b:</strong> CPA Experiments</td>
</tr>
<tr>
<td>Monday 28.11.2022</td>
<td>Meet Keti and report current plan and results</td>
</tr>
<tr>
<td>Monday, 12.12.2022</td>
<td>5\textsuperscript{th} Deliverable: Code and evaluation results</td>
</tr>
<tr>
<td></td>
<td><strong>Phase 1b:</strong> CPA improvements plan</td>
</tr>
<tr>
<td>Monday, 19.12.2022</td>
<td>6\textsuperscript{th} Deliverable: Intermediate presentation to Professor Bizer, code &amp; data</td>
</tr>
<tr>
<td></td>
<td><strong>Phase 2 (in 2 subgroups):</strong> Refinement</td>
</tr>
<tr>
<td>Monday, 16.01.2023</td>
<td>7\textsuperscript{th} Deliverable: Evaluation Results</td>
</tr>
<tr>
<td></td>
<td><strong>Phase 2 (in 2 subgroups):</strong> Error Analysis</td>
</tr>
</tbody>
</table>
## Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday, 16.01.2023</td>
<td>Meet Keti and report current plan and results</td>
</tr>
<tr>
<td>Monday, 06.02.2023</td>
<td><strong>8th Deliverable: Error Analysis Report</strong></td>
</tr>
<tr>
<td>Monday, 20.02.2023</td>
<td><strong>Phase 3: Transfer</strong></td>
</tr>
<tr>
<td>Monday, 06.03.2023</td>
<td>Meet with Keti and discuss plans and results</td>
</tr>
<tr>
<td>Monday, 06.03.2023</td>
<td><strong>Final Presentation to Professor Bizer</strong></td>
</tr>
<tr>
<td>Monday, 03.04.2023</td>
<td><strong>Phase 3: Write Report as HTML page</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Submission deadline</strong></td>
</tr>
</tbody>
</table>
Formal Requirements & Consultation

Deliverables

1. On the deliverable dates provide us via e-mail with:
   - Presentation slides
   - Task to member report: excel sheet stating which team member conducted which subtask
   - Code/Data: link or zipped folder with your code and data

2. Final Report as HTML-page
   - Giving overview of project results.
   - Providing artefacts/code, see team project

All deliverables should be sent to Keti & Chris!
Formal Requirements & Consultation

Final grade
- 20% for each phase (CTA, CPA, Refinement, and Transfer)
- 20% for final report
- Late submission: -0.3 per day

Consultation
- Send one e-mail per team or subgroup stating your questions to Keti
Useful Software

- **Transformers code**
  - Code to start in shared repository
  - HuggingFace Transformers: [https://huggingface.co/transformers/](https://huggingface.co/transformers/)
  - TURL: [https://github.com/sunlab-osu/TURL](https://github.com/sunlab-osu/TURL)
  - DODUO: [https://github.com/megagonlabs/doduo](https://github.com/megagonlabs/doduo)

- **Processing and GPUs**
  - Teaching GPU-Server
  - Google Colab: [https://colab.research.google.com/](https://colab.research.google.com/)
  - BwUniCluster2.0: [https://wiki.bwhpc.de/e/Category:BwUniCluster_2.0](https://wiki.bwhpc.de/e/Category:BwUniCluster_2.0)

- **Team Cooperation**
  - GitHub/Lab for the code base
  - Project Management Tool of your choice
Related Work: (Tabular) Transformers (1/2)


- D. Wang, P. Shiralkar, C. Lockard, B. Huang, X. L. Dong, and M. Jiang, “**TCN: Table Convolutional Network for Web Table Interpretation**,” arXiv:2102.09460 [cs], Feb. 2021
Related Work: (Tabular) Transformers (2/2)


Related Work: General (Deep) Schema Matching

Related Work: Table Annotation Benchmarks


- N. Abdelmageed, S. Schindler, B. König-Ries, **BiodivTab: A Table Annotation Benchmark based on Biodiversity Research Data**, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, volume 3103, pp. 13–18, 2021


- D. Ritze, C. Bizer, **Matching Web Tables To DBpedia - A Feature Utility Study**, in: Proceedings of the 20th International Conference on Extending Database Technology, pp. 210–221, 201

- More related work at PapersWithCode: [https://paperswithcode.com/task/column-type-annotation](https://paperswithcode.com/task/column-type-annotation), [https://paperswithcode.com/task/columns-property-annotation](https://paperswithcode.com/task/columns-property-annotation)
Questions?