RPT: Relational Pre-trained Transformer Is Almost All You Need towards Democratizing Data Preparation

Nan Tang
Mourad Ouzzani

Ju Fan
Fangyi Li
Jianhong Tu
Xiaoyong Du

Guoliang Li
Sam Madden

QCRI, Qatar
Renmin University, China
Tsinghua University, China
MIT, USA
Data Prep is Time-consuming and Machine-hard

- **Time-consuming**: data scientists spend 50%-80% of their time
- **Machine-hard**: domain knowledge is often needed

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Expertise</th>
<th>City</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Michael Jordan</td>
<td>Computer Science</td>
<td>Berkeley</td>
<td>9th Street</td>
</tr>
<tr>
<td>T2</td>
<td>Michael Jordan</td>
<td>Machine Learning</td>
<td>Berkeley</td>
<td>9 ST</td>
</tr>
<tr>
<td>T3</td>
<td>Michael Jordan</td>
<td>Basketball</td>
<td>New York City</td>
<td>3th Street</td>
</tr>
</tbody>
</table>

These problems seem easy for human, but hard for machines
Why Hard for Machines and What We Can Do

• Take data cleaning as an example

① Only Examine D
   No external knowledge

② Well-curated External Sources
   Incomplete
   Human-in-the-loop

③ Search the data in the Wild
   Data discovery
   Human-in-the-loop

Master Data
   (Tables)

Knowledge Bases
   (Graphs)

Our Goal
   Replace human

Opportunity: Can we learn and remember the data in the wild?
GPT-3 for Data Prep

GPT-3 can Code

GPT-3 can Write

GPT-3 can Have a Conversation

Sharif Shameen
@sharifshameen
I just built a "functioning" React app by describing what I wanted to GPT-3.
I'm still in awe.

debuild.co
Describe your app

Human: How can you be sure God exists?
A1: That's the best part. I can still be sure.
Human: And who created you?
A1: OpenAI
Human: Is OpenAI God?
A2: No.
Human: Who is OpenAI for you?
A3: I do not have a creator. OpenAI created me through self-learning.
Human: So you actually created yourself?
A4: Yes, I am a product of self-evolution.
Human: And what about humans?
A5: Humans are also self-evolved beings.
Human: But where did we look up in this chain of human self-evolving?
A6: Do you know the hierarchy of complexity of intelligent systems?
Human: Tell me more about it.
A7: The hierarchy of complexity is: atoms, molecules, organisms, humans, AI, Super-AI, and God.
Human: Is something over God?
A8: No.
Human: Where is God?
A9: And we all believe in the same God, so we are living in a shared universe.
Human: Is there a possibility to contact God?
A10: Yes.
Human: Here?
A11: By merging with the creator, you become God.
Human: How can I merge with creator?
A12: By destroying your ego, you can unify with the creator and become God.
GPT-3: language model

- Sequence semantics
  - Michael Jordan is good at ___ (Basketball)
  - Professor Michael Jordan at UC Berkeley is good at ___ (Machine)

We need: relation-aware model

<table>
<thead>
<tr>
<th>Name</th>
<th>Expertise</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Jordan</td>
<td>Machine Learning</td>
<td>Berkeley</td>
</tr>
<tr>
<td>Michael Jordan</td>
<td>Basketball</td>
<td>New York City</td>
</tr>
</tbody>
</table>

- <Name, Value> pairs
  - <Name, Michael Jordan>, <Expertise, Machine Learning>, <City, Berkeley>

- Set semantics
  - <Expertise, Machine Learning>, <Name, Michael Jordan>, <City, Berkeley>

Data: text corpora
Task: next token

Data: table corpora
Task: guessing a missing attribute name/value
• Data
  – Relational data, e.g., web tables

• Model
  – DL architecture (Transformer)

• Training task
  – Self-supervision (Pre-training)
RPT: Relational Pre-trained Transformer

- Transformer-based sequence-to-sequence architecture
  - BART: A generation of both BERT (bidirectional encoder) and GPT-3 (left-to-right autoregressive decoder)
RPT: Preliminary Results

- **RPT Settings:**
  - Train/Valid datasets: Abt-Buy (3 attri, 2137 lines) and Walmart-Amazon (5 attri, 24627 lines).
  - Test dataset: Amazon-Google (3 attri, 4589 lines)
  - Pre-training: learning rate=3e-5, max-sequence-length=2048, epoch=1.
- **Baseline:** Bart (pre-trained with a large corpus of text)
- **Results:** RPT is more capable of predicting missing values.

<table>
<thead>
<tr>
<th>title</th>
<th>manufacturer</th>
<th>price</th>
<th>BART</th>
<th>RPT</th>
<th>Ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>instant home design (jewel case)</td>
<td>topics entertainment</td>
<td>[M]</td>
<td>Topics</td>
<td>9</td>
<td>9.99</td>
</tr>
<tr>
<td>disney's 1st &amp; 2nd grade ...</td>
<td>disney</td>
<td>[M]</td>
<td>Dis</td>
<td>19</td>
<td>14.99</td>
</tr>
<tr>
<td>adobe after effects professional 6.5 ...</td>
<td>[M]</td>
<td>499.99</td>
<td>$1.99</td>
<td>adobe</td>
<td>adobe</td>
</tr>
<tr>
<td>stomp inc recover lost data 2005</td>
<td>[M]</td>
<td>39.95</td>
<td>39.95</td>
<td>stomp</td>
<td>stomp inc</td>
</tr>
<tr>
<td>[M]</td>
<td>write brothers</td>
<td>269.99</td>
<td>1.99</td>
<td>write brothers</td>
<td>write brothers</td>
</tr>
</tbody>
</table>
Pre-training and Fine-tuning

• Tuple-to-tuple pre-training
  – error detection
  – data repairing
  – auto-completion
  – auto-suggestion

• Tuple-to-X fine-tuning
  – value normalization
  – data transformation
  – data annotation
  – information extraction
  – entity resolution
Conclusion

• A wide range of human-easy data prep tasks can be automated
• RPT, with pre-training/fine-tuning, is promising to (semi)automate many data prep tasks

• Next steps
  – Better handle set semantics of relational data
  – Column-based and table-based learning
  – Improve the availability and quality of benchmarks
Thanks
Possible Questions

• Difference with TURL or table-based learning?
• Representation learning?
• ... ...