Extracting Zero-shot Structured Information from Form-like Documents: Pretraining with Keys and Triggers

Rongyu Cao, Ping Luo
Contents

• Background
• Challenge
• Solution
• Experiments
Background

Information Extraction from Document

• Input
  • a document (a collection of words with its content and position).
  • given attributes (keys).

• Output
  • for each key, which words consist the corresponding value.

Information Extraction from Plain Text

• Input
  • a sentence (a sequence of words with its content).
  • given attributes (keys).

• Output
  • for each key, which words consist the corresponding value.
Background

• Previous Method: LayoutLM [1]

• Framework
  1. Like BERT, take a collection of words with text embeddings and 2-d position embeddings as input.
  2. Like BERT, use multi-layers, multi-heads transformer.
  3. Predict a class label (multi-classification) for each word.
  4. The parameters are pre-trained on two tasks: masked visual-language model and multi-label document classification.

Challenge

• In real-world production, the keys needed to extract might have a huge number.

• Labeling large-scale training data for each key is costly.

• So, can the model generalize to unseen keys (zero-shot keys) without additional annotation?

• Thus, we should let the model learn key-invariant feature representation.
Solution

• What is the key-invariant feature representation?
• No matter what is the given key, the way we find value for the key is the same.
• Which way? How a human find the value for a given key?
Solution

- It explicitly learns two mappings: key-to-trigger and trigger-to-value.
How to annotate triggers and values for large-scale documents to pre-train this model?

Wikipedia Infobox [2]


Experiments

• Datasets

<table>
<thead>
<tr>
<th>description</th>
<th>access</th>
<th>#key</th>
<th>#zero-shot</th>
<th>#document</th>
<th>#pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Infobox Infobox on English Wikipedia webpage</td>
<td>crawler</td>
<td>7,370</td>
<td>-</td>
<td>398,467</td>
<td>1,238,281</td>
</tr>
<tr>
<td>Zh-Infobox Infobox on Chinese Wikipedia webpage</td>
<td>crawler</td>
<td>6,151</td>
<td>-</td>
<td>281,281</td>
<td>897,520</td>
</tr>
<tr>
<td>SROIE English receipts</td>
<td>public</td>
<td>6</td>
<td>2</td>
<td>972</td>
<td>5,505</td>
</tr>
<tr>
<td>Grater Chinese bank statements</td>
<td>private</td>
<td>24</td>
<td>8</td>
<td>4,032</td>
<td>18,825</td>
</tr>
</tbody>
</table>

Zero-shot keys for SROIE dataset: cash, telephone.

Zero-shot keys for Grater dataset: organization (开户机构), balance in previous page (上页余额), subject (科目), card number (卡号), balance (本页余额), unit (单位), print method (打印渠道), credit count (信用笔数).
Experiments

- Results

| row | method                      |  |  |  |  |  |  |  |  |  |  |
|-----|-----------------------------|---|---|---|---|---|---|---|---|---|
|     |                             | the SROIE dataset |  |  |  |  |  |  |  |  |  |
| 1   | BERT [Devlin et al. 2019]   | P  | R  | F1 | P  | R  | F1 | P  | R  | F1 |
| 2   | LayoutLM [Xu et al. 2020b]  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 3   | KA-bert                     | 0.1043 | 0.1000 | 0.1021 | 0.9373 | 0.9373 | 0.9373 | 0.2431 | 0.2422 | 0.2426 |
| 4   | KA-wiki                     | 0.6468 | 0.5961 | 0.6204 | 0.9460 | 0.9430 | 0.9449 | 0.3768 | 0.3702 | 0.3735 |
| 5   | KATA-bert                   | 0.0649 | 0.0471 | 0.0545 | 0.9366 | 0.9382 | 0.9347 | 0.3692 | 0.2491 | 0.2975 |
| 6   | KATA-wiki                   | **0.7678** | **0.6549** | **0.7069** | **0.9483** | **0.9457** | **0.9470** | **0.7378** | **0.7301** | **0.7339** |
| 7   | KATA-wiki w/ index embedding| 0.6569 | 0.6000 | 0.6283 | 0.9455 | 0.9423 | 0.9439 | 0.6809 | 0.6644 | 0.6725 |
| 8   | KATA-wiki w/o fine-tuning   | 0.5116 | 0.4314 | 0.4681 | 0.1106 | 0.0502 | 0.0691 | 0.1322 | 0.0796 | 0.0994 |
Conclusion

• We propose a novel task to extract the values of zero-shot keys in documents.

• We propose a two-stages framework to learn key-to-trigger and trigger-to-value mappings.

• We construct large-scale labeled dataset to pre-train the model on Wikipedia Infobox.
THANKS

2020.12.19