NAMED ENTITY RECOGNITION AND RELATION EXTRACTION WITH GRAPH NEURAL NETWORKS IN SEMI STRUCTURED DOCUMENTS

Manuel Carbonell, Pau Riba, Mauricio Villegas, Alicia Fornés and Josep Lladós

Computer Vision Center Barcelona
omni:us
Information extraction from semi-structured documents

GOAL: extract information from a document in a structured manner taking layout and semantics in account
Problem formulation

Information extraction can be reformulated as:

- **Word grouping**: aggregate words into entities
- **Entity labeling**: classify entities into categories (e.g. questions, answers and headers)
- **Entity linking**: find relationships between entities (possibly hierarchical)
Examples

In IEHHR entities are groups of words referring to a particular person in a marriage record. Links denote direct relationships between these persons e.g. wife - husband.

In funsd entities are groups of words that denote keys and values to be linked.
Methodology: Word graph

Node features := text box [ x, y, w, h, word_embed ]

GNN is trained for edge classification, to form word groups as connected components
Methodology: Entity graph

Node features := text box \[x, y, w, h, entity\_embed\]

GNN is trained for classification of nodes (entity labeling) and edges (entity linking)
## Results

<table>
<thead>
<tr>
<th></th>
<th>Word Grouping (ARI)</th>
<th>Entity Labeling (F1)</th>
<th>Entity Linking (F1)</th>
<th>External data</th>
<th># Params</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FUNSD [21]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[21]</td>
<td>0.41</td>
<td>0.57</td>
<td>0.04</td>
<td>✓</td>
<td>340M</td>
</tr>
<tr>
<td>[17]</td>
<td>-</td>
<td>0.79²</td>
<td>-</td>
<td>✓</td>
<td>160M</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>0.65</td>
<td>0.64</td>
<td>0.39</td>
<td>-</td>
<td>201M</td>
</tr>
</tbody>
</table>

| **IEHHR [22]**       |                      |                      |                    |               |         |
| **Ours**             | 0.65                | 0.53                 | 0.67               | -             | 201M    |
Results
Conclusions & future work

- GNN node and edge classification provides a promising method for entity recognition and relation extraction in semi structured documents.
- The proposed method has been designed for administrative documents but it can also be applied in other domains such as historical manuscripts.
- We believe that the obtained results have room for improvement and are limited due to the reduced size of the open available datasets for this type of task.
- Further research is required on a more larger openly available dataset for relation extraction and entity recognition in documents where semantic and spatial information plays a relevant role.