1. MOTIVATION

Worker with too many files to read

More and more common with the rate of data growth in the planet

But the gist of many professional / web documents is often in tables

2. RESEARCH GOAL

Automatically extracting information from tables in unstructured sources into databases OR tagging data in on-line tables with its relevant context will help workers cope with a flood of data AND will render data collection generally more efficient

3. PROCESS

A TABLE IS A GRID-LIKE STRUCTURE WHERE ITEMS ARE ALIGNED AND THIS GRAPHICAL COHERENCE CARRIES WITH IT A CONCEPTUAL COHERENCE

Quest for graphical coherence
First year and half of PHD

Steps 1-2: Location
For all lines, which are part of tables

Steps 3-4: Segmentation
For all table lines, group them into tables

Steps 5: Check 1
For all cells in a table, group them into columns

Quest for conceptual coherence
Second year and half of PHD

Steps 6: Function
Which columns and rows contain data or attributes?

Steps 7: Structure
Group each data field with its attributes

Steps 8: Interpretation/Check 2
Compare attributes against the knowledge base

As important as how each step performs individually is how they interact.

4. RESULTS SO FAR

<table>
<thead>
<tr>
<th></th>
<th>Complete- ness</th>
<th>Purity</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1+2</td>
<td>98%</td>
<td>2%</td>
<td>90%</td>
</tr>
<tr>
<td>Step 3</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Step 4</td>
<td>84%</td>
<td>90%</td>
<td>76%</td>
</tr>
</tbody>
</table>

5. SO WHAT’S NEW?

- Optimising the balance between tractability AND deferral of decisions to as late as possible by maintaining multiple hypotheses probabilistically and pruning out unlikely ones
- Dynamically integrating context knowledge representations to control the formation of the table and select between alternative models