Intelligent Indexing: A Semi-Automated, Trainable System for Field Labeling

Robert Clawson, Bill Barrett
Brigham Young University
What is indexing?
Our Research

- 5-10 times faster
- More accurate
- More enjoyable
Word Morphing

https://www.youtube.com/watch?v=eBQjHgejchA
Question

How can handwriting recognition be used to improve indexing?
Possible Methods

- Machine Learning Approach
- Split training/testset
- 1920 Utah Census
- About 50,000 fields per category
- Accuracy ~80%
  - Error rate too high
  - Would require lots of corrections after the fact
Possible Methods

- Pre-clustering
- Cost Matrix for each category/document

<table>
<thead>
<tr>
<th>Relationship to head of household</th>
<th>Daughter</th>
<th>Wife</th>
<th>Head</th>
<th>Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daughter</td>
<td>0</td>
<td>20.2</td>
<td>15.9</td>
<td>11.4</td>
</tr>
<tr>
<td>Wife</td>
<td>0</td>
<td>8.0</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>Head 1</td>
<td>0</td>
<td>0</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>Head 2</td>
<td>0</td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Possible Methods

- Pre-clustering
Possible Methods

- Pre-clustering

Problems

- Still makes frequent errors,
- Indexer has to scan up and down the page to look for mistakes
- How to show clusters when there are many different words in the column?
Breakthrough

- Interactive training: learn as you go, correct as you go
- Indexer drives
- Training set can start empty, quick ramp up
- Still use cost matrix
  - Switch from per document to per enumerator
- Introduce threshold
  - All fields that match under a threshold are labeled
  - Learn the threshold
- Demo
Training Set
Training Set
Training Set
You can help

- We need volunteers to test Intelligent Indexing

- To volunteer: email me (Robert Clawson) at: intelligentindexing@gmail.com