Predicting Role Relevance with Minimal Domain Expertise in a Financial Domain

Mayank Kejriwal
"For the year ended December 31, 2015, billed business from charge cards comprised 56 percent of total U.S. Card Services billed business. Centurion Bank and American Express Bank as Issuers of Certain Cards and Deposit Products
We have two U.S. bank subsidiaries, American Express Centurion Bank (?Centurion Bank?) and American Express Bank, FSB (?American Express Bank?), which are both FDIC-insured depository institutions. Certain information regarding each bank is set forth in the table below:"
Domain Expertise: what do we mean?

• Definitions of terms (taken from NYS-CPA)

**Accrued Interest**: INTEREST that has accumulated between the most recent payment and the sale of a BOND or other fixed-income security.

• Definitions of roles

**GUARANTOR**: A legal arrangement involving a promise by a person (guarantor) to perform the obligations of a second person (or many persons), in the event that the latter person fails to meet their obligations.

• Characteristics of the problem to be solved
• Feature crafting
• Ability to diagnose unexpected outputs
• Many more!
Motivation 1: Feature crafting is tricky!
Motivation 2: Unlabeled data can be powerful...if used right

• Intuitively, a source of (additional) background knowledge
• What if there’s no human?
Motivation 3: Impressive recent advances in NLP

- Mainly due to neural networks and low-dimensional latent space modeling
Case study: *2vec (word2vec, doc2vec...)

Unsupervised learning method for representing text and non-textual data in a vector space.
Approach

• Use labeled+unlabeled text data to train a vector for each ‘contextual fragment’ (after some preprocessing)

"For the year ended December 31, 2015, billed business from charge cards comprised 56 percent of total U.S. Card Services billed business. **Centurion Bank** and **American Express Bank** as Issuers of Certain Cards and Deposit Products. We have two U.S. bank subsidiaries, **American Express Centurion Bank** (?Centurion Bank?) and **American Express Bank, FSB** (?American Express Bank?), which are both FDIC-insured depository institutions. Certain information regarding each bank is set forth in the table below:"
Train role-specific classifiers

• Use explicitly given role-specific relevance labels (ignore documents without labels in training)
## Training results

- Consistent with (eventually released) ground truth 2 results

<table>
<thead>
<tr>
<th>Role</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>affiliate</td>
<td>0.941176471</td>
<td>0.941176471</td>
<td>0.941176471</td>
</tr>
<tr>
<td>trustee</td>
<td>0.976744186</td>
<td>1</td>
<td>0.988235294</td>
</tr>
<tr>
<td>issuer</td>
<td>0.909090909</td>
<td>0.909090909</td>
<td>0.909090909</td>
</tr>
</tbody>
</table>
Conclusion

• Extremely simple to set up (1-2 hours of programming and set up)
• Will likely improve with more data, whether labeled or unlabeled
• Does not require us to understand the domain
• Could also work in tandem with hand-crafted features
• Achieved median results on all ground truths