KBC from richly formatted text using Fonduer: Applications to financial reports

Andrea Aguirre, Roberto Seminario

CS230: Final Report

Motivation

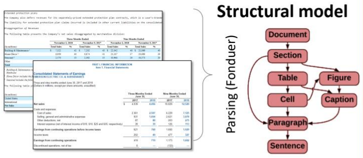
- Our objective: create a model that learns to **extract revenue information** from richly formatted financial 10-Q filings
- Complicated task because attributes and relations are expressed in a combination of textual, structural, tabular and visual signals

Dataset

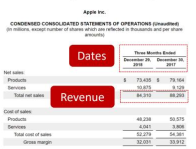
Get Data from SEC

- PostgreSQL
- 2,007 reports
- HTML format
- Text spans: +60 million

Structural model



Defining training candidates



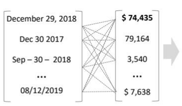
- **Candidate**
  – (Date, Revenue) pair
- **Potential # candidates**
  – 60 mm * 60 mm

Total: 3.5e+15

Too many so we filter

Hard-filtering – limit # of candidates

- Logical, tabular, format, content, linguistic, RegeX rules to limit # of candidates



Total: 17,217
Training examples

Weak supervision

- Manual labeling is unfeasible, so we use data programming (Snorkel)
- **Labeling functions (LFs)** evaluate the relation between the mentions of each candidate

Example labeling functions (we have 14 in total)

```
def LF_vertical_align(c):
    (period, revenue) = c
    per = period.context.sentence
    rev = revenue.context.sentence
    if per.col_start <= rev.col_start and per.col_end >= rev.col_end: return TRUE
    else: return ABSTAIN

def LF_early_in_table(c):
    (period, revenue) = c
    per = period.context.sentence
    rev = revenue.context.sentence
    if per.row_start <= 3 and rev.row_start <= 12: return TRUE
    else: return ABSTAIN
```

GAN for label generation

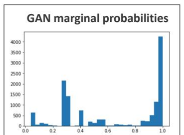
- Apply LFs to unlabeled data, resulting in a label matrix Λ. Then encode generative model pw(Λ, Y) using three factor types: labeling propensity, accuracy, and pairwise correlations of LFs

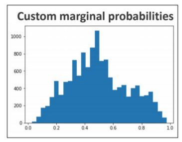
$$\phi^{\text{Lab}}_{i,j}(\Lambda, Y) = \mathbb{1}\{\Lambda_{i,j} \neq \emptyset\} \qquad p_w(\Lambda, Y) = Z_w^{-1} \exp\left(\sum_{i=1}^{m} w^T \phi_i(\Lambda, y_i)\right)$$

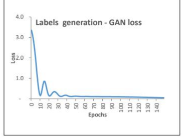
$$\phi^{\text{Acc}}_{i,j}(\Lambda, Y) = \mathbb{1}\{\Lambda_{i,j} = y_i\}$$

$$\phi^{\text{Corr}}_{i,j,k}(\Lambda, Y) = \mathbb{1}\{\Lambda_{i,j} = \Lambda_{i,k}\} \qquad \hat{w} = \arg\min_w - \log \sum_Y p_w(\Lambda, Y)$$

Creating weak labels







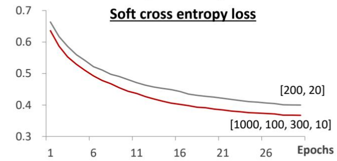
Will try both training marginals to choose the best performing one

Feature matrix

- **107,000 features:** One-hot vectors of surrounding words, style features (font, size, bold, caps, etc.), NLP structure, length, lemma sequences, row and column headers, table characteristics, page location, object hierarchy, html tags, and others

Results

- **Chosen model**: Shallow [1000, 100, 300, 10] neural network with a soft cross entropy loss function using Adam optimizer



larger NN

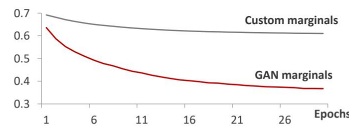
	Development	Test
Accuracy	33.8%	35.7%
Precision	24.8%	28.0%
Recall	95.3%	87.8%
F1 score	39.4%	42.4%

Smaller NN

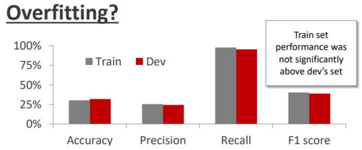
	Development	Test
Accuracy	27.2%	31.0%
Precision	23.2%	27.2%
Recall	96.3%	92.8%
F1 score	37.4%	42.0%

[Performance vs true labels - obtained from Capital IQ]

Choice of marginals



Overfitting?



Conclusions

- Through a combination of hard filters and weak supervision, our model was able to pinpoint a handful of revenue + date candidates out of billions of potential pairs
- F1 performance was slightly below 40%, with a baseline performance of 0% for the broad dataset and 20% for the filtered one
- Although performance was decent, more than half of the model's performance was in the hard-filtering portion (20%).

Expansions

- Use fewer hard filters and train model with more candidates (over 1mm ideally)
- To avoid computation unfeasibility, trim features from 107,000 to less than 10,000
- Overall, creating a structural model is computationally inefficient for a task like this. Table and page extraction based on heuristics paired with a machine learning model would be more efficient

References

[1] Sen Wu, Luke Hsiao, Xiao Cheng, Braden Hancock, Theodoros Rekatsinas, Philip Levis, and Christopher Ré. 2018. Fonduer: Knowledge Base Construction from Richly Formatted Data. In Proceedings of 2018 International Conference on Management of Data, Houston, TX, USA, June 10–15, 2018 (SIGMOD'18), 16 pages.

[2] Alexander Ratner, S. H. (November 2017). Snorkel: Rapid Training Data Creation with Weak Supervision. Stanford University.

Team members

- Andrea Aguirre, andreaar@stanford.edu
- Roberto Seminario, rseminar@stanford.edu