Towards Automatic Extraction of References Between Text and Tables

Sherry Ruan
Joint work with Juho Kim and Maneesh Agrawala

Stanford University

March 10, 2016

Overview

1. Big Vision
2. Related Work
3. Our Work
   - Data Collection
   - Auto-Generated UI Demo
   - Automatic Classification
   - Interactive Reading System
4. Challenges

Big Vision

- Analyze how information is referenced at scale
- Automatically extract references between text and charts
- Use extracted references to build a better interactive paper reading user interface

Related Work

- Revision [Savva et al. 2011]

Related Work

- Revision [Savva et al. 2011]

  $\surd$ Identify the chart type using computer vision and machine learning

Related Work

- Revision [Savva et al. 2011]
  - $\surd$ Identify the chart type using computer vision and machine learning
  - $\surd$ Extract the graphical marks and infer the underlying data

Related Work

- Revision [Savva et al. 2011]
  - $\surd$ Identify the chart type using computer vision and machine learning
  - $\surd$ Extract the graphical marks and infer the underlying data
  - $\surd$ Apply cognitive design principles to improve graphical perception

Related Work

- Revision [Savva et al. 2011]
  - $\surd$ Identify the chart type using computer vision and machine learning
  - $\surd$ Extract the graphical marks and infer the underlying data
  - $\surd$ Apply cognitive design principles to improve graphical perception
  - $\times$ Cannot relate chart data to text data

Related Work

- Revision [Savva et al. 2011]
  - $\surd$ Identify the chart type using computer vision and machine learning
  - $\surd$ Extract the graphical marks and infer the underlying data
  - $\surd$ Apply cognitive design principles to improve graphical perception
  - $\times$ Cannot relate chart data to text data
- Extracting References via Crowdsourcing [Kong et al. 2014]

Related Work

- Revision [Savva et al. 2011]
  - √ Identify the chart type using computer vision and machine learning
  - √ Extract the graphical marks and infer the underlying data
  - √ Apply cognitive design principles to improve graphical perception
  - × Cannot relate chart data to text data
- Extracting References via Crowdsourcing [Kong et al. 2014]
  - √ Present a crowdsourcing pipeline to extract the references

Related Work

- Revision [Savva et al. 2011]
  - √ Identify the chart type using computer vision and machine learning
  - √ Extract the graphical marks and infer the underlying data
  - √ Apply cognitive design principles to improve graphical perception
  - × Cannot relate chart data to text data
- Extracting References via Crowdsourcing [Kong et al. 2014]
  - √ Present a crowdsourcing pipeline to extract the references
  - √ Provide an interactive document viewing application

Related Work

- Revision [Savva et al. 2011]
  - √ Identify the chart type using computer vision and machine learning
  - √ Extract the graphical marks and infer the underlying data
  - √ Apply cognitive design principles to improve graphical perception
  - × Cannot relate chart data to text data
- Extracting References via Crowdsourcing [Kong et al. 2014]
  - √ Present a crowdsourcing pipeline to extract the references
  - √ Provide an interactive document viewing application
  - × Crowd algorithm is not scalable

Related Work

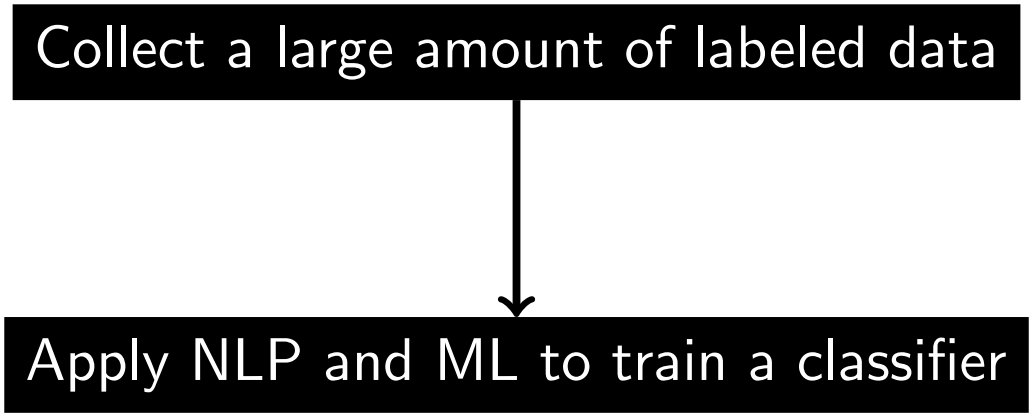
- Revision [Savva et al. 2011]
  - √ Identify the chart type using computer vision and machine learning
  - √ Extract the graphical marks and infer the underlying data
  - √ Apply cognitive design principles to improve graphical perception
  - × Cannot relate chart data to text data
- Extracting References via Crowdsourcing [Kong et al. 2014]
  - √ Present a crowdsourcing pipeline to extract the references
  - √ Provide an interactive document viewing application
  - × Crowd algorithm is not scalable

Our Goal: Automatically extract references at scale!

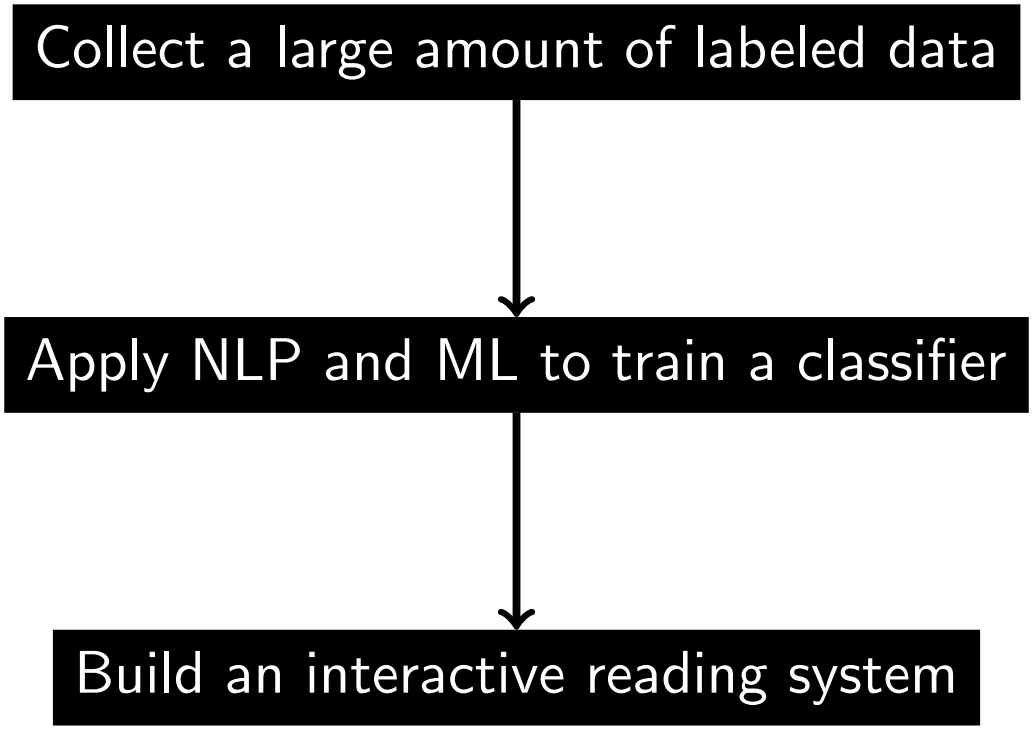
Overview of Our Work

Collect a large amount of labeled data

Overview of Our Work



Overview of Our Work

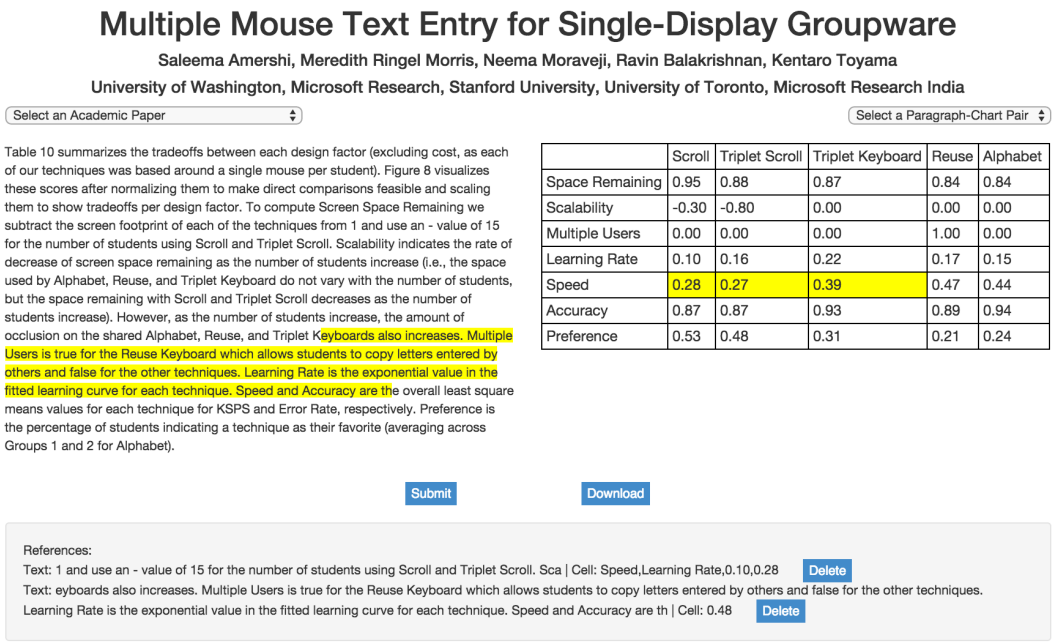


Data Collection from Crowd

- Automatically extract tables from pdf files (tools available: Tabula, pdffigures)
- Pair each extracted table with its corresponding paragraph
- Automatically generate easy-to-use websites to collect references from crowd
- Currently focus on text-cell annotation

Auto-Generated UI Demo

- Extract tables from pdf files by analyzing their surroundings
- Represent extracted tables as bounding boxes in a coordinate system
- Automatically generate a web page for every pair of extracted tables and pre-selected paragraphs



Multiple Mouse Text Entry for Single-Display Groupware

Saleema Amershi, Meredith Ringel Morris, Neema Moraveji, Ravin Balakrishnan, Kentaro Toyama

University of Washington, Microsoft Research, Stanford University, University of Toronto, Microsoft Research India

Select an Academic Paper

Select a Paragraph-Chart Pair

Table 10 summarizes the tradeoffs between each design factor (excluding cost, as each of our techniques was based around a single mouse per student). Figure 8 visualizes these scores after normalizing them to make direct comparisons feasible and scaling them to show tradeoffs per design factor. To compute Screen Space Remaining we subtract the screen footprint of each of the techniques from 1 and use an - value of 15 for the number of students using Scroll and Triplet Scroll. Scalability indicates the rate of decrease of screen space remaining as the number of students increase (i.e., the space used by Alphabet, Reuse, and Triplet Keyboard do not vary with the number of students, but the space remaining with Scroll and Triplet Scroll decreases as the number of students increase). However, as the number of students increase, the amount of occlusion on the shared Alphabet, Reuse, and Triplet Keyboards also increases. Multiple Users is true for the Reuse Keyboard which allows students to copy letters entered by others and false for the other techniques. Learning Rate is the exponential value in the fitted learning curve for each technique. Speed and Accuracy are the overall least square means values for each technique for KSPS and Error Rate, respectively. Preference is the percentage of students indicating a technique as their favorite (averaging across Groups 1 and 2 for Alphabet).

	Scroll	Triplet Scroll	Triplet Keyboard	Reuse	Alphabet
Space Remaining	0.95	0.88	0.87	0.84	0.84
Scalability	-0.30	-0.80	0.00	0.00	0.00
Multiple Users	0.00	0.00	0.00	1.00	0.00
Learning Rate	0.10	0.16	0.22	0.17	0.15
Speed	0.28	0.27	0.39	0.47	0.44
Accuracy	0.87	0.87	0.93	0.89	0.94
Preference	0.53	0.48	0.31	0.21	0.24

Submit Download

References:

Text: 1 and use an - value of 15 for the number of students using Scroll and Triplet Scroll. Sca | Cell: Speed,Learning Rate,0.10,0.28 Delete

Text: eyboards also increases. Multiple Users is true for the Reuse Keyboard which allows students to copy letters entered by others and false for the other techniques. Learning Rate is the exponential value in the fitted learning curve for each technique. Speed and Accuracy are th | Cell: 0.48 Delete

Automatic Classification

- Baseline algorithm uses exact word match
- Attempt different preprocessing methods: normalizing, stemming and lemmatization
- Results largely depend on papers given

Interactive Reading System

- Construct intermediate representations of tables
- Establish a mapping between extracted tables and original pdf files via a coordinate system
- Highlight related cells when selecting phrases in papers, and vice versa

Challenges

- Implement tools that can extract complex tables (i.e., spanning multiple rows, without borders)
- Analyze semantic meanings of words to associate them with data in tables
- Generalize to different types of charts

References

Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala and Jeffrey Heer (2011)

ReVision: Automated Classification, Analysis and Redesign of Chart Images

User Interface Software and Technology (UIST), Oct 2011, pp. 393-402.

Nicholas Kong, Marti A. Hearst and Maneesh Agrawala (2014)

Extracting References Between Text and Charts Via Crowdsourcing

ACM Human Factors in Computing Systems (CHI), Apr 2014, pp. 31-40.

Thank You!