Extracting Names Using Layout Clues: An Initial Report

Aaron P. Stewart and David W. Embley
Department of Computer Science, Brigham Young University

Abstract
Successfully extracting and indexing names from OCRed historical documents is particularly challenging, and precision and recall are low compared to extracting names in similar clean text. We note, however, that many historical documents contain formatted lists containing names; these lists have rich visual context, but valuable layout clues are not passed on or used in name recognizers. We show in this paper how to exploit these clues to improve precision and recall results for OCRed historical documents that contain list patterns for names. Our solution automatically analyzes a document to identify name-list patterns and applies them to better identify names in OCRed documents. Preliminary results show that the application of name-list patterns improves recall by finding names that could not have been recognized by even the best dictionary-based recognizers and improves precision by eliminating non-name tokens inadvertently labeled as names by recognizers that are too permissive.

1 Introduction

Historical books such as city directories and family histories contain a wealth of genealogical information. It is now feasible for libraries to digitize these books and make the content searchable. Typically, the visual digitization is good, but the OCR is often poor, making searchable text of lower quality. Furthermore, we would like to do more than just search for text strings. Ideally, we would like to extract names, dates, and locations and relationships among them and make these extracted genealogical facts directly queriable. In addition, we would like to provide provenance information for query results, so that clicking on a result would display an image of the original document with the information in the results highlighted so that a user could see and check the results extracted from original documents.

As a first step toward this goal, we would like to be able to extract names of people. The names may be hard to identify for several reasons: (1) OCR errors, which make the tokens output by the OCR engine unreadable even though a human could easily read the name in the image. (2) Some names are the same as ordinary words in the text; we should not extract these words as names. (3) Some names may not appear in name dictionaries, but should still be extracted as names. (4) Some names appear as parts of something else such as a company name, a school name, or a street name and should ot be extracted as person names.

Typical named entity recognizers examine text and text context to determine where names begin and end. [Finkel05]. Often they use simple dictionary lookup. They may also look at internal character n-grams. These internal features help to identify names that have never been seen before, or that are corrupted by noise. These recognizers rely on machine learning techniques and are typically trained to work with running text, where names appear as part of a complete sentence. They usually perform poorly on OCRed text and on formatted text such as name lists.

A second approach involves hand crafted rules and dictionaries [Grover08]. The Ontology Extraction System (OntoES) also takes this approach to named entity recognition [Embley99]. We use OntoES as our baseline extractor in our work.

These named entity recognizers, however, typically do not take advantage of formatting information. Considering Figure 1, which is a city directory, it is easy to see a list pattern for names. A name appears at the beginning of a non-indented line. A comma terminates the name on the right. Following the comma additional information appears, sometimes with text that looks like names but is not a name. The list-of-names recognizers we present in this paper identify names in lists by left and right context, which may be either spatial or textual.

When names truly are in lists, this technique improves recall because it identifies names in the list that may not be identified by other means. For example, in Figure 1, the baseline recognizer does not identify "Coleman Ebenezer" or "Coleman Rodey" because these given names are not in the lexicon. The technique improves precision because it eliminates spurious names that are in the text but not in the list. For example, in Figure 1, the baseline recognizer identifies "Myrtle" and "T Loo", but these should not be recognized as names. The first, which comes from the end of the line of the last entry ("Coleman Taylor") is a location designator. The other comes from an OCR error at the bottom of the page, in which the "2300" and part of the "M" from "C. M. SMITH" are recognized by the OCR engine as "T Loo". These errors are visible in bold in Figure 2, which contains the beginning and end of the OCR text for the page in Figure 1.

We present the details or our list-pattern recognizer and some preliminary results showing how it improves both recall and precision as follows. In Section 2, we describe our baseline name extractor. In Section 3, we give a brief overview of the preprocessing steps. Section 4 describes the algorithm for identifying tab stops. In Section 5, we explain how to apply the results in our system, and in Section 6 we give some qualitative results.

ADAMS & FALLS. Real Estate & Insurance, 2005 1st AVENUE,

BIRMINGHAM CITY DIRECTORY. 185

Cole Wm C (Reamer, Cole & Co), res 411 19th
Cole Wm D, boarding, 19th cor Arlington ave, Bessemer
Cole Wm P, res 1712 2d ave
Coleman Adaline, c, domestic J L Davis
Coleman Bettie, domestic E K Fulton
Coleman C C, vice-pres and mngr Birmingham Tool and Implement Works, res 2417 6th ave
Coleman Daniel, wks H A & B R R
Coleman David, c, coachman J D Dwyer, bds 12th ave s w cor 17th s s
Coleman Ebenezer, c, wks H A & B R R, res 2107 ave C
Coleman Frank, lab L & N R R
Coleman Frank H, clk Whilden & Campbell, res 1504 4th ave
Coleman George, c, lab Alice Furnace
Coleman George, lab, bds n s 3d ave 1 e of 12th
Coleman Granville L, agt The Bradstreet Co, rooms 22 Roden block
Coleman G W, clk L & N R R, bds 19th bet 4th and 5th aves
Coleman Henry, foreman L & N R R, res 223 ave C

W. S. Smith & Co., Architects. Plans, Specifications, And Superintending,

Coleman Henry, c, lab Alice Furnace, bds e s 17th 5 n of ave I s s
Coleman H, pudler Bham Rolling Mill
Coleman Ida, c, domestic, Morris Avenue Hotel
Coleman James G, clk D R Copeland & Co, res 411 22d
Coleman John, wks H A & B R R
Coleman John W, carpenter, C T Hughes & Co, res 16th near 15th ave
Coleman Lewis, c, wks Sloss Furnace, res 2d al bet 24th and 25th s s
Coleman Ludie, c, domestic H M Archibald
Coleman Malachi, c, lab, bds s s 1st al bet aves A and B
Coleman Patrick, brickmason, bds 19th bet 4th and 5th aves, Bessemer
Coleman Prescilla, c, cook F B Hemphill, bds 2105½ 2d ave
Coleman Rodey, lab L & N R R
Coleman Ruffin, physician, 2019½ 2d ave, room Roden Bldg
Coleman R E, brakeman L & N R R
Coleman Seaborn, c, res al bet 25th and 26th n of 5th ave
Coleman Taylor, c, lab, res al bet 1st and 2d aves e of Myrtle, Avondale

C. M. SMITH & CO. 2300 MORRIS AVENUE. LUMBER DEALERS. Special Selections for FLOORING, CEILING & FINISHING

TELEPHONE 132 For Everything in its Season, THE BEST AT ALL SEASONS. T I T C H E 213, 22d STREET.

Figure 1. A sample page from a city directory.

```
« i ADAVS & FA..S. Real Estate
&Insnrance, 2006 1st AVENUE,
BIRMINGHAM CITY DIRECTORY. IST)
Cole Wm C (Reamer, Cole & Co),
res 411 19th
...
Coleman Taylor, e, lab, res al
bet 1st and 2d aves e of Myrtle,
Avondale I"* III OLIITIJ JP Pfl
LUMBER DEALERS. LJ. 'T Loo . Ofl
Ifl Ot !lAI. Special Selection,
for M^R.6 AVENUE. FLOORING,
CEILING & FINISHING H m r m CO?
"I m f H H£ B/2 H<£ wsr" MC wSS
HCR ¥ £5 00 f t g$ g* CD O&2 va>
CCJD <3D O 1 H to t » "h to T £
Si Q » w w 3« • • '
```

Figure 2. Selected OCR text from the sample page.

2 Baseline Name Extractor

Our baseline name extractor is independent of the visual layout and formatting of a page. It accepts input from the OCR engine, such as the data shown in Figure 2. It uses hand-coded regular expressions and dictionaries to identify names of people.

The baseline extractor is not infallible. Our algorithm is based on the idea that the baseline is good enough to help identify patterns, and that the patterns will give an improvement over the baseline.

3 OCR Preparation

Our corpus consists of images, text, and word bounding boxes from a third-party source. Because the data comes from a variety of OCR engines, we run some preprocessing steps to put the data in a suitable form.

A first step is separating columns and pages. In our example, we have manually cropped one page from a two-page image. Cropping prevents subsequent algorithms from interleaving pages or columns.

Next, the code combines bounding boxes into lines. Although the OCR engine has already identified lines, we repeat the process using our own algorithm to ensure uniformity.

Figure 3. Word bounding boxes.

4 Marker Insertion

Now that the page is separated into lines, we need to find the tab stops. OCR is inherently a noisy process. Page rotation and other process variations make it unreasonable to expect lines to be exactly left-aligned.

In order to find tab stops, we use a modified RANSAC algorithm [Fischler81]. The basic idea is to select a pair of bounding boxes and draw an imaginary stripe connecting the left corners of the two bounding boxes and extending across the length of the page. If a maximal number of bounding boxes are within a threshold of this stripe, then the stripe is accepted as a tab stop. The affected lines of text are marked accordingly, and the process is repeated on the remaining lines of text.

In Figure 4, the lines indicate the relevant tab stops that were found.

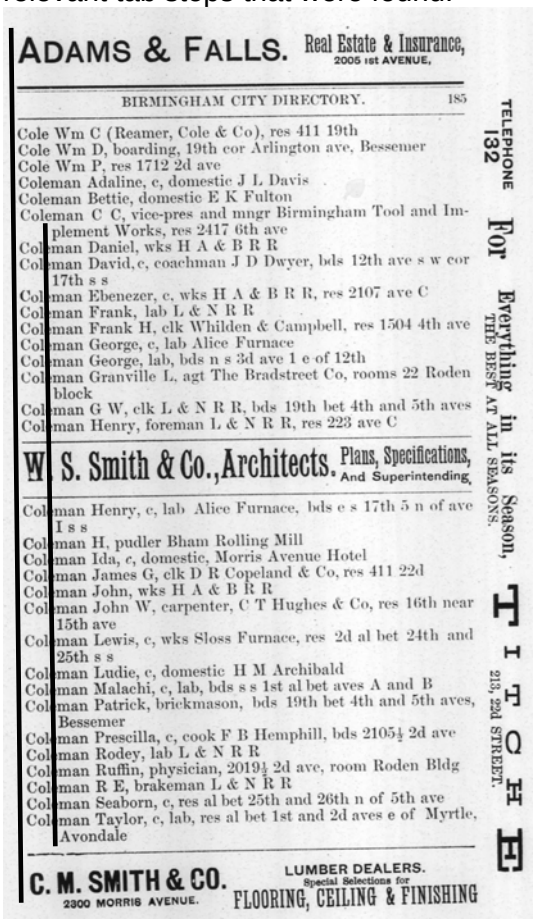

Figure 4. Tab stops identified by the modified RANSAC algorithm.

Once the tab stops are found, the code inserts markers into the token stream.

```
[NO-TAB]« [NO-TAB]i [TAB1]ADAVS
& FA..S. Real Estate &Insnrance,
[NO-TAB]2006 1st AVENUE,
[TAB3]BIRMINGHAM CITY DIRECTORY.
[NO-TAB]IST) [TAB1]Cole Wm C
(Reamer, Cole & Co), res 411
19th [TAB1]Cole Win D, boarding,
19th cor Arlington ave. Bessemer
[TAB1]Cole Wm P, res 1712 2d ave
[TAB1]Coleman Adaline, c,
domestic J L Davis [TAB1]
```

Figure 5. Token stream with formatting markers.

5 Pattern Finding and Name Extraction

Once the baseline extractor has identified names and the RANSAC margin finder has identified tab stops, the algorithm searches for salient patterns. The process is simple.

For each name identified by the baseline, the algorithm takes the token immediately to the left and the token immediately to the right. These left/right pairs are counted and sorted. Any token pairs with a count above a threshold count are taken as potential patterns. The pattern with the highest count is taken as the most salient pattern.

```
Left context: [TAB1]
Right context: ","
21 occurrences
```

Figure 6. The most salient context pair from the example.

Now that the patterns are identified, the algorithm applies the most salient pattern to extract names from the document. The extractor simply looks for strings surrounded by the left and right context, and extracts them as names. Results are given in the following section.

6 Sample Results

We are still working towards an annotated corpus in which we can appropriately and completely test our system. However, we will give preliminary results of running the algorithm on this particular example, in Figure 7 and Figure 8.

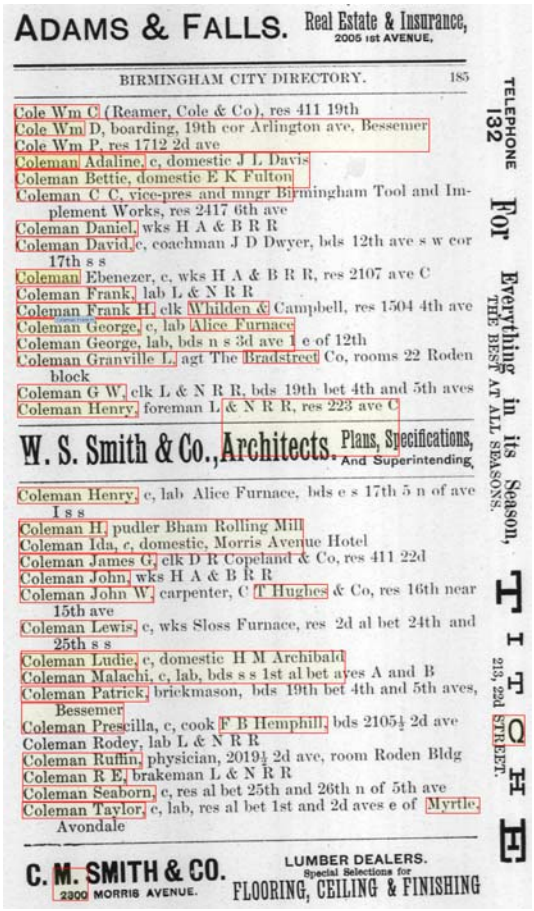

Figure 7. Baseline name extraction. *Large rectangles are artifacts of the display code and result from phrases that span multiple lines. This image has been edited to compensate for a display error.*

The pattern-based extractor improves recall by correctly recognizing newlines. For example, the baseline extractor (Figure 7) extracts incorrect names such as “Bessemer Cole Wm”, “Davis Coleman Bettie”, and “Mill Coleman”. These cost both precision and recall because the incorrect match

overlaps the true value and prevents it from being found. The pattern-based extractor correctly identifies these regions.

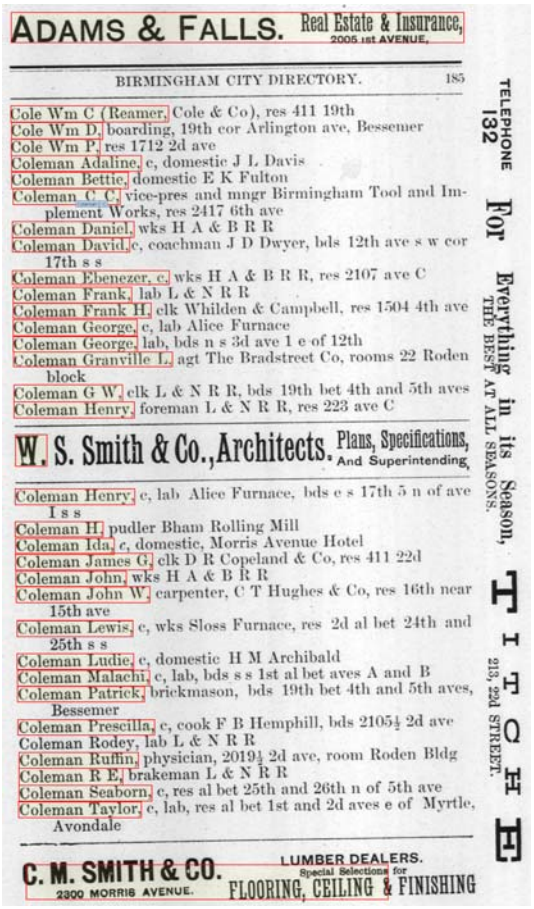

Figure 8. Pattern-based extraction.

The pattern-based method also improves precision by not picking up spurious entries in the middle of the text. The baseline extracted the street name "Myrtle", the company name fragments "Bradstreet" and "Whilden &", and some misrecognized OCR text near the middle of the page. The pattern extractor did not pick up these entries.

The pattern extractor still leaves room for improvement. For example, it does not pick up the names "H M Archibald" or "F B Hemphill" from the inner regions of the text. It can also run past the end of a name if the right context is absent or lost in the OCR process. In Figure 8, the first name is misrecognized as "Cole Wm C (Reamer".

Overall, the results give an encouraging picture of the possibilities of pattern-based extraction. We believe our approach is useful when applied appropriately.

On our selected example, the baseline precision was 56.1%, with a recall of 62.16% and an F1-score of 58.97%. The pattern approach gave a precision of 82.35%, a recall of 75.68%, and an F1-score of 78.87%. These figures should not be taken as authoritative, since this is a hand-picked training example.

7 Conclusions

We have presented a simple system that may improve name recognition in certain types of documents. We expect to find further improvements as we apply the system to a larger volume and variety of data.

For future work, we need to develop or find an annotated corpus containing an ample quantity and variety of data within this domain. We expect that additional development data will drive further improvements in the algorithm. A blind data set from a separate set of books within the genre would enable a meaningful evaluation of the results.

We may extend the pattern finder to work with centered and right-justified names, and other text patterns.

We also may implement better handling for the right context of the patterns. When the right context is a comma and the OCR software misses it, other heuristics could help determine where the name ends.

A practical system should also discriminate between regions that contain helpful patterns and regions that do not.

While much work remains, our preliminary results suggest that formatting information can be used to improve name recognition.

8 Acknowledgements

This project was supported in part by Ancestry.com. We are grateful for their contribution.

9 References

[Embley99] D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith, Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages, Data & Knowledge Engineering, 31(3):227–251, November, 1999.

[Finkel05] J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

[Fischler81] M.A. Fischler and R.C. Bolles (June 1981). "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography". Comm. Of the ACM 24: 381–395. doi:10.1145/358669.358692.

[Grover08] C. Grover, S. Givon, R. Tobin and J. Ball, Named Entity Recognition for Digitised Historical Texts, Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, May, 2008.

[Packer10] T.L. Packer et al., Extracting Person Names from Noisy OCR Text. Unpublished manuscript, December 2010.

[Smith09] Smith, Ray. Hybrid Page Layout Analysis via Tab-Stop Detection. Proceedings of the 10th international conference on document analysis and recognition, 2009.