Extracting Person Names from Diverse and Noisy OCR Text

Thomas Packer, Joshua Lutes, Aaron Stewart,
David Embley, Eric Ringger, Kevin Seppi

Department of Computer Science
Brigham Young University
Provo, Utah, USA

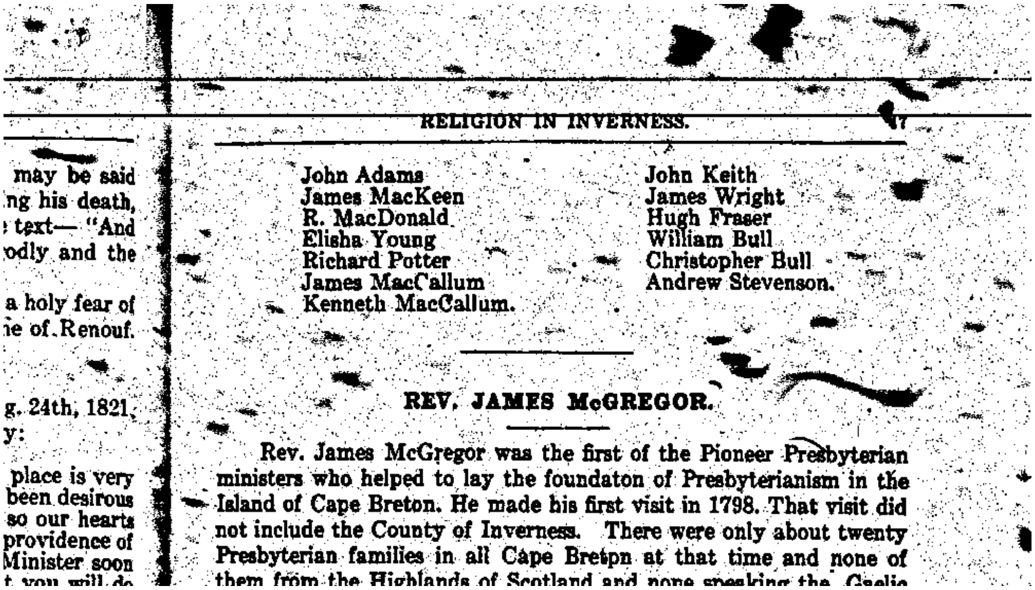
Lee Jensen

Ancestry.com, Inc.
Provo, Utah, USA

Information extraction can be of great value to genealogy and other historical research. The more automatically and accurately we can categorize the words appearing in the scanned images of a document, the more value a researcher can get out of a large collection of documents. For example, the greater the number of person names that can be correctly identified in a collection (the higher the recall), the more likely those names will be discovered by someone searching for them. Likewise, the fewer the number of words that are incorrectly categorized as being part of a person's name (the higher the precision), the less time it will take a researcher to sift through a collection to find information about a person of interest.

In the research presented here, we were concerned with extracting person names from scanned document images, including books and newspapers. In doing this, we faced two challenges. The first was extracting information from a collection of documents with great genre and format diversity and with few specialized resources from which to develop the extractors (e.g. labeled training data for machine learning approaches). Pages of our test set were sampled from 10 documents from the following genres: family and local history books, newspapers, city directories, Navy cruise books and church congregation year books.

The second challenge was extracting information from the noisy text produced by the processes of scanning and OCR transcription. In the dataset[i] we used, these processes introduce the following kinds of errors: character substitutions; incorrectly split or joined words; and word-order, text-line boundaries and other document structure information being lost by post-processing of the OCR output. For example, in the section of the page shown in Figure 1, the two-column list of names runs together with text on the opposite page to produce a non-delimited stream of words. Splotches on the page also cause character recognition errors.



may be said
ng his death,
text— "And
odly and the

a holy fear of
ie of Renouf.

g. 24th, 1821.
y:

place is very
been desirous
so our hearts
providence of
Minister soon
t you will do

RELIGION IN INVERNESS.

John Adams
James MacKeen
R. MacDonald.
Elisha Young
Richard Potter
James MacCallum
Kenneth MacCallum.

John Keith
James Wright
Hugh Fraser
William Bull
Christopher Bull
Andrew Stevenson.

REV. JAMES McGREGOR.

Rev. James McGregor was the first of the Pioneer Presbyterian ministers who helped to lay the foundaton of Presbyterianism in the Island of Cape Breton. He made his first visit in 1798. That visit did not include the County of Inverness. There were only about twenty Presbyterian families in all Cape Bretpn at that time and none of them from the Highlands of Scotland and none speaking the Gaelic

Figure 1. The upper right-hand corner of one of the "noisier" pages in our collection. The OCR transcription of the first three lines read as follows: "ot Bo quo qu 4 o ot 4 uo q o qu uot q i JKIF q o t q o ot quot q u i I q John Adams John Keit h to have died with his armour on The his death James MacIceen James Wright though infirm and ill he Sunday preached preceding to his ople from the text - R MacDonald Huh Fraser"

We applied the following six techniques for extracting person names. Their extraction qualities are compared in the graph in Figure 2. The lighter green bar (left side of each pair) is the exact-match F-measure where an extracted name is considered to be correct only if every word within that name is classified as part of the name and no additional words are joined with the name. the darker blue bar (right side of each pair) is the fragment-match F-measure where credit for correct and incorrect classification is given on a word-by-word basis. In both cases, F-measure is the harmonic mean of precision and recall. These metrics are computed against the hand-annotated full names in our test set.

- ***Dictionary***. Words found in substantial dictionaries of given names, initials, surnames and titles were labeled as such. Contiguous sequences of labeled words were extracted as full names.
- **Regular Expression (*Regex*)**. Similar to *Dictionary*, but only full names that matched one of several hand-written regular expression patterns were output, e.g. "Title GivenName Initial Surname".
- **Context-free Grammar (*CFG*).** Similar to *Regex* but with a more complicated set of hand-written rules in a more expressive weighted context-free grammar.
- **Maximum Entropy Markov Model (*MEMM*)**. This statistical sequence model was trained on hand-labeled newswire text from a CoNLL dataset[ii]. The trained model was then applied to the noisy OCR text with some domain adaptation attempted.
- **Conditional Random Field (*CRF*)**. Another well known statistical sequence model trained and executed just like the *MEMM*.
- ***Ensemble***. The best ensemble among several variations was one that allowed the Dictionary, Regex and MEMM extractors to vote. It output a full name when two or more agreed.

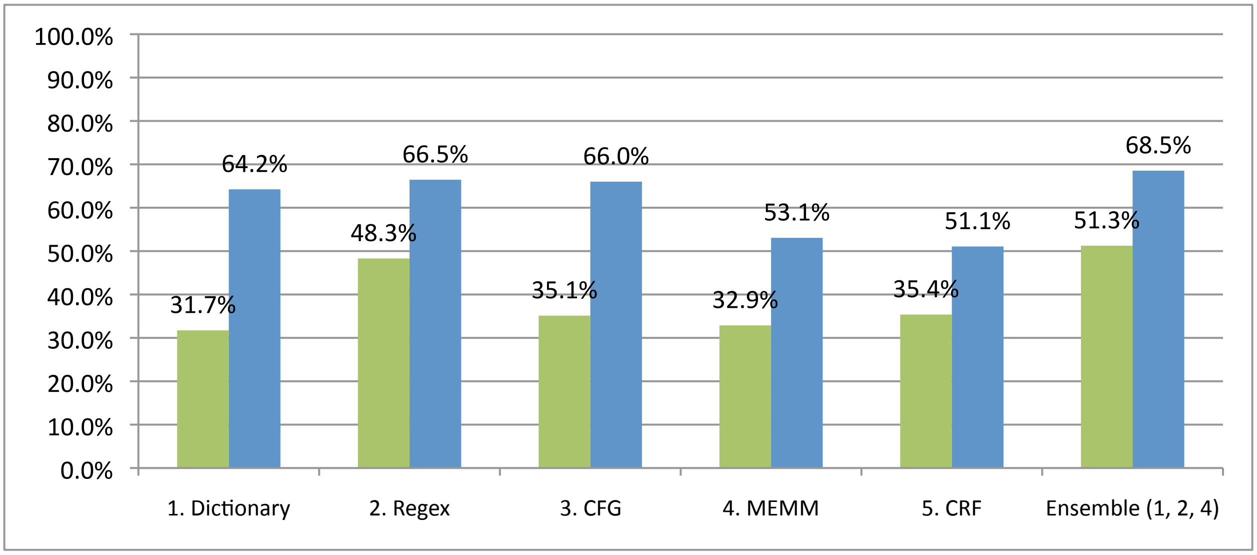


Figure 2. Exact-match (left) and fragment-match (right) F-measure of each method evaluated on all 10 documents of the test set. There were one to two pages of hand annotated text taken from each document and placed in this test set.

In conclusion, from this project we learned that relatively simple methods (1 and 2) can outperform more sophisticated methods in a limited-resource setting. We were also encouraged to see an improvement over the individual methods by using a simple voting ensemble.

There is still room for improvement. We will take the results of, and experience gained from, this initial project as a starting point on which to build better, more adaptive extractors using page format recognition, OCR error correction and other advanced techniques.

[i] Thanks to Ancestry.com, Inc. for the use of this data.
[ii] http://acl.ldc.upenn.edu/W/W03/W03-0419.pdf