PatientBank: Medical Record Text Extraction with Dictionary-based Error Correction

Mikayla Thompson

December 17, 2014

1 Abstract

PatientBank currently offers medical record request and online storage services for patients and their doctors. It would be helpful for both to be able to have text searching capabilities, but the medical record documents are often stored as scanned image files, which contain no textual information. Tesseract is an optical character recognition engine from Google that can extract text from documents, however the results it produced were quite messy and ineffective. Dictionary-based error correction was developed to select likely-erroneous words, and search for alternatives within a short edit-distance. This error correction improved the document coverage by up to 60.3%. However, in very visual images with complicating artifacts, even robustly corrected error rates are 62.2%, which is lower than the target rate. The most promising technique for improving this would be targeting the error correction at optically likely errors, rather than a generalized search.
2 Background: PatientBank

PatientBank is working to provide individuals with a way to gather their own health information. Medical records, especially for patients with complicated histories, are often split between many healthcare providers and are difficult to request and collect. PatientBank works with the patient and providers to request, compile, and maintain a comprehensive record on the patient’s behalf.

3 Introduction

PatientBank’s current model centers on document storage and sharing: patients request their medical records from previous healthcare providers, they’re provided in electronic form or scanned in, and then patients can share the documents with their current providers. Both patients and doctors are sifting through large amounts of information and search capabilities would be very helpful in finding clinically important documents. However, many documents have been scanned from paper copies and are in effect image files and not amenable to text searches. This project (a) utilized optical character recognition (OCR) technology to extract searchable text from document images; and (b) found likely-inaccurate OCR artifacts and corrected them to English words (essentially a spell check).

4 Methodology

This program is written as a Ruby 2.1.2 command line script, and also packaged as a component of a gem.
4.1 Extract

The fundamental step in going from image to text is the optical character recognition. The OCR engine used in this project is Tesseract[1]. It was originally developed at HP labs, and is now maintained by Google.

4.1.1 Implementation

My contribution in this area is essentially a Ruby wrapper that calls Tesseract from the command line. The constraint of the tesseract program is that it will only accept as input an image filename and a text filename to output to. It won’t write to stdout, for instance, so the wrapper needs to work around these constraints.

An ImageMagick Image object is used to initialize a text_box object. If a filename is provided instead, the image is loaded.

When the get_text method is invoked, the image is saved to a temporary file, and the Tesseract command line is called with the image file and a temporary text file. When that program finishes, the text file is loaded and sanitized (stripped of non-letter/digit characters, lowercased, whitespace condensed) for output.

4.1.2 Results

The quality of the output depends enormously on the properties of the input image. For fairly regular, typeset documents, the results are quite accurate. The sample in Figure 1 has 11 erroneous characters of 262 total, for an error rate of 4.2%.

Looking at the errors that are made, several of them are single character errors within a word (note ”yeaxs” for ”years”, and ”tubetculin” for ”tuberculin”. It appears
possible that a simple spelling correction algorithm would be able to find the intended word and resolve these errors.

4.2 Correct

The second stage is to apply a spelling correction algorithm. Given the subject matter, it was necessary to consider medical as well as everyday vocabulary, and it is also necessary to allow for proper nouns that may not be in a dictionary (patient names, doctor names, etc.). The output text will be used for searches, so it is preferable to have false positives (an irrelevant document is offered) than false negatives (a relevant document is not offered). The appropriate compromise is to institute spell checking (with English and medical dictionaries) but to add the most likely candidates to the document without replacing the target word. In the event that the target word is a name, for instance, it will not be deleted from the search text.
4.2.1 Implementation


The spelling correction has a default edit distance of 1, but the option of increasing it to 2. These options are compared in the Results section.

After the document text has been extracted, each word (splitting on spaces and punctuation) is checked against the combined dictionary. If no entry is found, all words with an edit distance of one/two from the original word are calculated. Any words of this candidate set that are found in the dictionaries are added to the text document, along with the original word.

This algorithm is loosely based on the spelling corrector famously written by Peter Norvig in 20 lines of Python [4]. The most substantial difference is that Norvig’s algorithm is trained on a corpus of language in use, rather than a dictionary. He makes the assumption that the most-used of the candidate words is the most likely to be the intended word. This is helpful in the case where an incorrect word is being replaced by a correct word. However, for this project, it’s unnecessary to insert a single word, and practical to instead include all of the possible terms within a single edit.

4.2.2 Results

Figure 2 shows the proposed alternatives for the non-dictionary words in Figure 1, with an edit distance of 1.

Clearly some of the errors (2, 6) were corrected appropriately. Others were not.

From this data and larger samples, it can be shown that each uncorrected error was not resolved for one of three reasons:
1. The intended word (from the image) is not in the dictionary.

2. The output word (from the OCR algorithm) is not the intended word, but is also in the dictionary.

3. The edit distance between the intended and output word is greater than the defined limit.

In error 3, 7, and 8, the edit distance is too far too be corrected (Reason 3). Error 4 comes Reason 1, the intended word is not in the dictionary. Error 1 is an interesting special case of Reasons 1. The word from the image ("diptheria") is not in the dictionary (Reason 1), so the algorithm searches for an alternative. It finds one, "diphtheria", which is the more standard spelling. Because both versions will be in the text, this document can be found by searching for either spelling.

An example of Reason 2 would be if the OCR engine read "hyperglycemic" as "hypoglycemic", also a valid word, but a very different one. The algorithm would not enter the alternative finding stage, because it has no way of knowing that "hypoglycemic" was not the intended word.
5 Results

A more complete test on two visit summaries was conducted and quantified to compare the results with and without spelling correction, as well as between different forms of documents. Both documents are attached to this report.

"Master" refers to a manually transcribed text copy to use as a comparison. "Output" is the result of applying the OCR engine, and, in the second and third rows, spelling correction. The "Overlapping" column is the number of words that the two documents have in common. Accuracy measures how many of the words found in the output document are actually in the master document. Less accuracy increases the chance of false positives: the document showing up in searches where it does not contain the search terms. Coverage is the most relevant variable, and the one being optimized for in this case. It is the percent of words in the master document that are also found in the output. 100% coverage would mean that searching for any word in the master would bring up the document in a search.

The first document was particularly challenging because it used multiple typefaces, variable spacing between words and lines, decorative structuring (boxes around particular sections) and had a irregular background. OCR alone was able to produce 38.8% of the words in the file.

<table>
<thead>
<tr>
<th>Type</th>
<th>Master Words</th>
<th>Output Words</th>
<th>Overlapping</th>
<th>Accuracy</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Output</td>
<td>201</td>
<td>312</td>
<td>78</td>
<td>25.0%</td>
<td>38.8%</td>
</tr>
<tr>
<td>Corrected (1 edit)</td>
<td>201</td>
<td>1059</td>
<td>103</td>
<td>9.7%</td>
<td>51.2%</td>
</tr>
<tr>
<td>Corrected (2 edits)</td>
<td>201</td>
<td>6567</td>
<td>125</td>
<td>1.9%</td>
<td>62.2%</td>
</tr>
</tbody>
</table>
Notably, of the 76 words not found by the 2 edit correction, 29 of them are not in either of the dictionaries, which means that the correction algorithm found 72.6% of "find-able" edits.

The second document was more well-formed: an direct-to-pdf export from an electronic medical system that was never scanned.

<table>
<thead>
<tr>
<th>Type</th>
<th>Master Words</th>
<th>Output Words</th>
<th>Overlapping</th>
<th>Accuracy</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Output</td>
<td>104</td>
<td>106</td>
<td>102</td>
<td>96.2%</td>
<td>98.1%</td>
</tr>
<tr>
<td>Corrected (1 edit)</td>
<td>104</td>
<td>211</td>
<td>102</td>
<td>48.3%</td>
<td>98.1%</td>
</tr>
<tr>
<td>Corrected (2 edits)</td>
<td>104</td>
<td>1584</td>
<td>103</td>
<td>6.5%</td>
<td>99.0%</td>
</tr>
</tbody>
</table>

As expected, the scores are much, much higher, with nearly perfect coverage.

We expect that the medical records we receive will span a wide range in terms of quality and ease of integration, so it

6 Conclusion

The mechanisms described—using the Tesseract OCR engine, coupled with spelling correction using English and medical dictionaries—appear to be very promising for indexing the documents for search. Well-formed documents achieve coverage rates of 98% from OCR alone, but with a more complicated, scanned document, implementing a spelling correction algorithm increases the coverage by 60%.

In the future, several components could be added to increase this coverage rate. One of the forms of edit that is not currently included is combinations of neighboring words. Often, a word is split (e.g. "emer ency"), giving two words that are both a
significant edit distance away from the original word. The algorithm in its current state does not try to combine words, so an error like this would remain uncorrected.

Additionally, a more nuanced interpretation of errors and their sources would be helpful in suggesting edits. For instance, ”g” to ”a” is a common error in the output from the OCR engine. Instead of trying all possible edits, it would make sense to try combinations of likely edits. ”Emetanfg” is 3 edits away from ”emergency”, but each of the edits is an optically logical one, and similar for ”otqoic” to ”atopic” or the 7th error in Figure 2. Focusing on optically likely changes will allow exploration to a larger edit distance in those directions without requiring an exponentially increasing number of candidates.

The long term plan is that Patient Bank will not simply store documents, but extract the medical information from them, and use it to build a single, integrated medical record for the patient. In order to get there, it’s necessary to be able to very reliable get information out of the document, and understand the structure that it came from. This implementation of OCR with error correction will be applied as a useful step towards that goal.

References

Holly 4903  female  July 1, 2008

prepared for July 28, 2009
age 1 year 18 days
last visit April 28, 2009
next visit July 27, 2009
for 1YR HGB,PB,HIB,HBV,M

PROBLEMS
1 CRADLE CAP (11-7-08)
2 ATOPIC DERMATITIS (11-7-08)

TREATMENTS
1 EMOLLIENTS/WESCORT OINT LIGHTLY BID X 4.6 D (1-13-09)

ALLERGIES
NKDA

IMMUNIZATIONS
DPT series
DTaP 1 9/10/08 2m
DTaP 2 11/7/08 3m 27d
DTaP 3 1/13/09 6m 3d

HIB series
PRPOMP 1 9/10/08 2m
PRP T 2 11/7/08 3m 27d
PRP T 3 1/13/09 6m 3d

POLIO series
IPV 1 9/10/08 2m
IPV 2 11/7/08 3m 27d

PEUMO series
PNCV7 1 9/10/08 2m
PNCV7 2 11/7/08 3m 27d
PNCV7 3 1/13/09 6m 3d

ROTAVIRUS series
ROTAS 1 9/10/08 2m
ROTAS 2 11/7/08 3m 27d
ROTAS 3 1/13/09 6m 3d

MRRV series
MMR series

INFLUENZA series
most recent 2/17/09 7m 7d
previous 1/13/09 6m 3d

HEPATITIS series
HepB ped 1 9/10/08 2m
HepB ped 2 11/7/08 3m 27d

PAPILLOMA series
MENING series

TYPHOID series
BCG series

misc immunizations ...
contraindications ...

SCREENINGS
last TB (Mantoux) test: result
last hemoglobin: result
last blood lead: result
last urinalysis: result
last crisis episode: cumulative count:
last UTI episode: cumulative count:

PERCENTILES
ht: 74 cm (29.1 in) 86% ile on 4/28/09 (8m 18d)
OFC: 46 cm 92% ile on 4/28/09 (8m 18d)

body surface area: 0.46 sq M
body mass index: 18 kg/m sq
predicted peak flow:

last WCC visit: 4/28/09 at age 9m 18d
next allowed WCC visit: 7/28/09 at age 1y 18d

SPECIAL INFORMATION
reactive airway history
hemoglobinopathy
cardiac history
major allergy history
metabolic problem
seizure history

callback for referate: restrictions:

RECOMMENDATIONS
MEASLES-MUMPS-RUBELLA (MMR) vaccine (live) is suggested for this patient.
VARICELLA vaccine (live) is suggested for this patient.
HEPATITIS A vaccine (any variety: 0.5 ml) is suggested for this patient.
PNEUMOCOCCAL CONJUGATE (Prevenar) vaccine is suggested for this patient.
A scheduled HEMOGLOBIN TEST is suggested for this patient.

*
### Patient Demographics

<table>
<thead>
<tr>
<th>Patient Address</th>
<th>Communication</th>
<th>Language</th>
<th>Race / Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Home</td>
<td>English (Preferred)</td>
<td>White / Not Hispanic or Latino</td>
</tr>
<tr>
<td></td>
<td>Emergency Contact</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mobile</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Note from UW Health - Wisconsin

This document contains information that was shared with [Redacted]. It may not contain the entire record from UW Health - Wisconsin.

### Reason for Visit

<table>
<thead>
<tr>
<th>Reason</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knee Problem</td>
<td>New left knee fibular pain</td>
</tr>
</tbody>
</table>

### Encounter Details

<table>
<thead>
<tr>
<th>Date</th>
<th>Type</th>
<th>Department</th>
<th>Care Team</th>
</tr>
</thead>
<tbody>
<tr>
<td>08/13/2013</td>
<td>Office Visit</td>
<td>RESEARCH PARK SPORTS MEDICINE</td>
<td>Scerpella, Tamara A, MD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>621 SCIENCE DR</td>
<td>621 SCIENCE DR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MADISON, WI 53711</td>
<td>MADISON, WI 53711</td>
</tr>
<tr>
<td></td>
<td></td>
<td>608-263-8850</td>
<td>263-8850</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>265-8340 (Fax)</td>
</tr>
</tbody>
</table>

### Active Allergies - as of 12/06/2014

No Known Allergies

### Medications - as of 12/06/2014

<table>
<thead>
<tr>
<th>Prescription</th>
<th>Sig.</th>
<th>Disp.</th>
<th>Refills</th>
<th>Start Date</th>
<th>End Date</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>norgestimate-ethinyl estradiol (ORTHO-CYCLEN) 0.25-35 MG-MCG per tab</td>
<td>one time daily.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Active</td>
</tr>
</tbody>
</table>

### Active Problems - as of 12/06/2014

<table>
<thead>
<tr>
<th>Problem</th>
<th>Noted Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posterior tibial plateau fracture</td>
<td>02/19/2013</td>
</tr>
</tbody>
</table>