Evaluation of Handwriting Recognition Systems for Application to Historical Records

Patrick Schone, Heath Nielson, Mark Ward
{patrickjohn.schone,NielsonHE,WardRM}@ldschurch.org
FamilySearch, 50 E North Temple, Salt Lake City, UT

ABSTRACT

In the last decade, significant, largely-governmental funding has been applied to the automatic transcription of handwritten documents. Uses for this kind of technology are somewhat limited given that the numbers of handwritten documents are on the decline. However, certain types of handwritten historical records can be crucial for genealogical research in that they identify key vital facts. In recent years, organizations like FamilySearch have exhausted huge efforts to identify, digitize, and transcribe these kinds of genealogically-rich records. Until now, such transcription has largely been done through massive crowd-sourced labor. We believe handwriting recognition technology is only a few years away from profitable application to genealogical documents. To test this hypothesis, we developed an evaluation paradigm for measuring handwriting recognition performance on four data collections of differing genres and languages. We invited research organizations to participate in the evaluation and compared performance to the outcome of human annotation. In this paper, we provide the details of this paradigm, including the guidelines, corpora and evaluation tools. Then we illustrate the exciting system results which suggest that the state-of-the-art is very close to providing real-world benefit to the automatic transcription of genealogically-rich documents.

1. BACKGROUND

Vital, census, legal, migration and other types of historical records have significant value for genealogical research. This kind of information can be crucial for identifying key facts in the lives of ancestors. In recent years, organizations such as FamilySearch have exhausted significant effort to identify, digitize, and transcribe these kinds of genealogically-rich records. The bulk of the transcription of these digitized records has largely been done through massive crowd-sourced labor.

There are several concerns about a human transcription process. First, transcription can be arduous which makes it difficult to attract large cadre of well-trained volunteers. Furthermore, there is a substantial need for transcription of foreign documents and unusually-formatted record types, but for these kinds of materials, there are very limited pools of qualified or interested transcribers. Lastly, since there is no particular guarantee of speed nor accuracy, it becomes necessary to identify individuals who can recruit, train, and motivate the sometimes-volunteer transcription workforce and to validate the quality of their work. The ability to automate this transcription process could have huge payoff.

Over the last decade, there have been sizeable monetary and technical contributions to the automatic transcription of offline (i.e., previously-written) handwritten documents. This process is often referred to as *handwriting recognition* (HR). The DARPA MADCAT program [1], for instance, provided a huge infusion of funding into HR which resulted in rapid improvements in this technology. Additionally, the NIST OpenHART [2] evaluation allowed research organizations with HR systems to compare their algorithms on common data sets. Thus, whereas the accuracies of HR systems in the early parts of the 2000's were negligible, the system with the highest reported score a decade later at OpenHart 2010 had a word error rate of 37.7% (or 62.3% accuracy). Although this accuracy was based on an assumption that word boundaries would be provided, this level of accuracy still suggests that HR has advanced to the point of likely being usable in a production environment.

Could such technologies be applied to historical documents of genealogical value? The major emphasis of previous HR funding programs and evaluations had been for free-flowing, unstructured documents in particular languages of interest to funding sponsors. So if HR systems could be usable for genealogical records, they would need to be redeployed from operating on free-form document formats to documents that are typically structured or semi-structured. They would likewise need to automatically identify regions of interest for transcription and would need to be rapidly portable into new languages and environments for which they were not originally designed.

To determine if current systems could be retargeted in these ways, FamilySearch created a major evaluation (referred to here as *IRIS*) consisting of four genealogically-relevant document types spanning three different languages. The tools and collections assembled for this massive evaluation consisted of tens of thousands of historical images for training and testing; corresponding but imperfect transcripts; some English linguistic materials; and a highly flexible scoring algorithm. No bounding boxes were provided, nor were there any additional non-English resources. Human performance was also evaluated for the major English collection.

Using the training materials, interested system builders were given 90 days to train their systems. They were then provided with the test data and given a two-week window to apply their algorithms to it. Systems were then scored with a weighted word error rate (WWER), which favored information of higher genealogical value (such as personal name components). A number of research organizations with HR systems were invited to participate, and several received copies of the training data, but only two were able to provide results to the competition, namely A2iA (from France) and BBN (from Boston).

Given the OpenHart results, human error rates, the difficulty of the algorithm-porting task, the lack of word boundaries, the cross-lingual issues, the noisy transcripts used for training, and the weighting on the harder-to-transcribe elements, it was expected that system WWER would be no better than 40-50% in English and potentially much worse in other languages. In fact, one of the collections was so difficult, we expected performance thereon to not break 100% WWER. Much to our surprise, however, systems were able to achieve WWERs as low as 19.6%! Moreover, even on the extremely difficult collection, where no gains were expected and humans even struggled to provide benefit, HR systems were able to get as low as 92.4% WWER.

We believe that the outcome of this evaluation suggests that HR systems are just at the door of being able to provide significant benefit to genealogical record transcription. With limited additional funding in this domain, there could be a huge acceleration in the indexing of genealogically-relevant historical documents.

In this paper, we describe this exciting evaluation more fully. We provide documentation about the collections and metrics, and we show the best results on each collection as contrasted with some human performance. We also invite other interested parties to participate in this evaluation.

2. EVALUATION COLLECTIONS

The IRIS evaluation consisted of four different collections of data which were selected to study system performance along various axes related to document format and language. The particular collections were: (a) the 1930 US Census, (b) the 1930 Mexico Census, (c) Arkansas marriage records, and (d) French Parish records. Table 1 indicates the size of the training and evaluation corpora that were prepared for each collection. The evaluation data were held back from system developers until the competition time, but developers were told that it largely matched the training data structure (with the limited exception being that for censuses, where the evaluation would include records from a state that was not in training).

Corpus	Training Size	Evaluation Size
1930 US Census	15,061	1,673
1930 Mexico Census	8,652	961
Arkansas Marriages	7,502	834
French Parish Records	10,529	1,170

Table 1: Numbers of Training Documents Per Collection

2.1 United States 1930 Census Collections

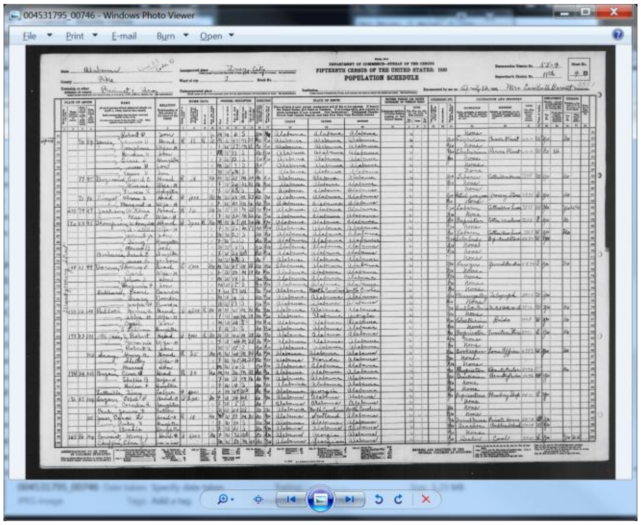

004531795_00746 - Windows Photo Viewer
File
Print
E-mail
Burn
Open
DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
FIFTEENTH CENSUS OF THE UNITED STATES: 1930
POPULATION SCHEDULE

Figure 1A. United States 1930 Census

The US census records (shown in Figure 1A) were selected because they are tabular and assumed to be the easiest to transcribe. Moreover, since censuses are considered to be the most genealogically-beneficial document collections, automatic transcription of them could have huge genealogical payoff.

At the time of the IRIS preparation and evaluation (June 2011-March 2012), the 1940 Census had not yet been released and the 1930 Census had not yet been fully transcribed. Yet other censuses, such as the 1920 census, were available. So in addition to image and transcripts that were provided for training materials, statistics on 1920 personal given names and surnames were also provided for possible use in developing language models.

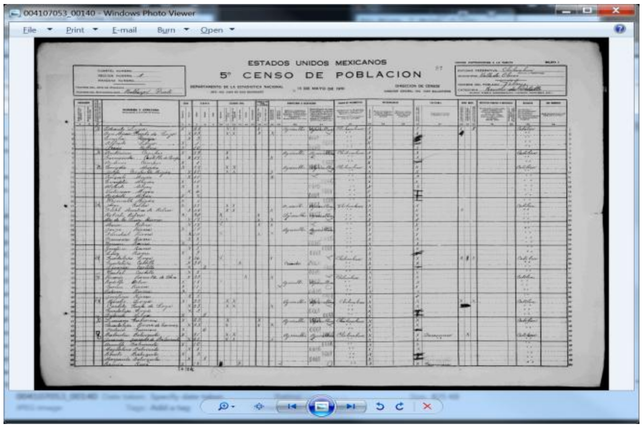
The training transcripts that were provided for this as well as all other collections were created by hosts of volunteer annotators. As such, they contain errors. Moreover, the transcriptions were prepared for other purposes independent of the IRIS evaluation, so there were conventions that were followed which resulted in non-verbatim transcriptions. For example, a person from the census who was born in Pennsylvania may have a birth place listed as Pennsylvania, PA, Penn, Penn., etc., or by some ditto information (DO or ‘’). Yet annotators would have been instructed to record each of these as “Pennsylvania.” This phenomenon occurred regularly and required special attention during system building as well as in the creation of scoring tools.

In addition to these issues of the transcripts being *inexact*, the transcripts were also *incomplete*. The formulaic page and column headers and footers were not transcribed, so systems would have to ensure that this information did not show up in their outputs. Likewise, not all columns of data were transcribed, so results from these would also need to be overlooked by the systems.

In the case of the U.S. Census only, the *evaluation* data was re-transcribed by a commercial transcription company that was tasked with attaining at least 99.5% transcription accuracy. This high accuracy was sought for IRIS (as well as for a separate, coincident project) so that any observed errors in the final evaluation – at least on this collection – would be almost exclusively due to HR system issues.

Before leaving the description of the U.S. Census, there is one more key point that needs to be mentioned. An additional confounding issue for system-builders for the US Census and other data sets was that IRIS provided no bounding boxes or human-provided clues to where the key genealogical facts could be observed on each page. So systems needed to automatically detect the layout of each page, and identify each column, row, and box. For some participants, this became a major focus since it was not a requirement for the free-flowing documents with which they had developed significant experience.

2.2 Mexico 1930 Census Collections



ESTADOS UNIDOS MEXICANOS

5° CENSO DE POBLACION

Figure 1B. Mexico 1930 Census

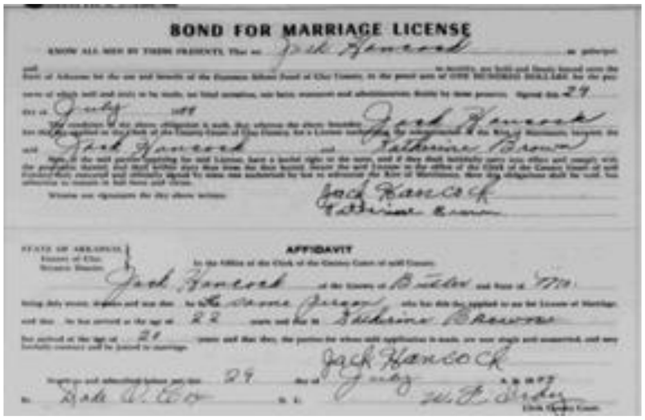
If automatic transcription has value, it needs to be able to be reconfigurable to new languages with limited effort. To test this language portability, the Mexico 1930 Census (Figure 1B) was also chosen, which is fully in Spanish.

Additionally, in the evaluation, system builders were restricted from augmenting their training algorithms with information that might give them an advantage which would not be available to system users. For example, it would be unfair to transcribe documents from the Rhode Island 1930 census with a foreknowledge of all of the names of the 1930 census. Though it was hard for the IRIS evaluation creators to explicitly control for this kind of contamination, it seemed that incorporating the Mexico census might help since it was likely that participants would have had limited previous exposure to that collection.

Another benefit of using this collection is that no other transcribed Mexico censuses existed at the time of the evaluation. Thus, utilizing HR on this collection may result in a good representation of system performance on a completely new collection type with limited or no former word-usage statistics for boosting results.

There are also some idiosyncracies with this data. The marriage information in this collection is not provided as text, but is provided as columnar checkboxes. The census-taker was instructed to mark an "X" in the *Soltero* column if the individual was single, an "X" in the *Casado Por Lo Civil* column for civilly married, and so forth. This meant that the HR systems would positively have to identify the correct column in order to accurately transcribe the data.

2.3 Arkansas Marriage Collections



BOND FOR MARRIAGE LICENSE

Jack Hancock

Jack Hancock

Jack Hancock

Katherine Brown

Jack Hancock

Katherine Brown

AFFIDAVIT

Jack Hancock

22

Katherine Brown

Jack Hancock

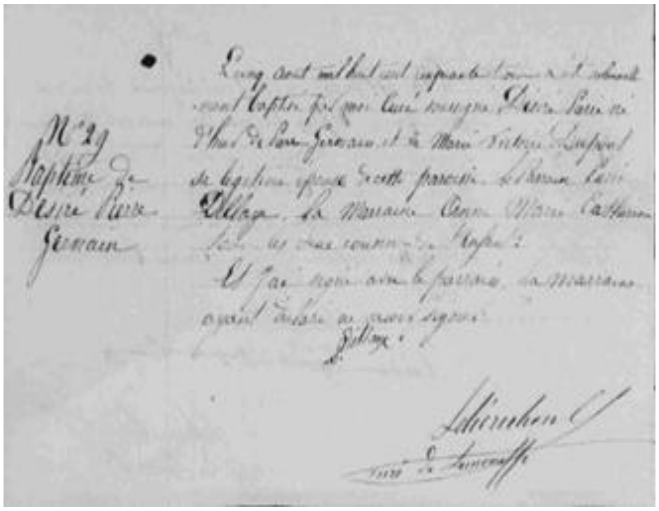
Figure 1C. Arkansas Marriages

Not all genealogical collections are in tabular form. Others are fill-in-the-blank templates, such as the Arkansas Marriage collection (see Figure 1C). The key genealogical pieces of information identified on these records include the groom and the bride, the dates and places, and any available extended family information. The rest of the printed and handwritten components of the page are not provided in the training transcripts nor of interest in the evaluation. Thus, like the other collections, a system needs to automatically identify the specific regions of genealogical interest and transcribe only those portions.

However, there are some benefits to this kind of collection. Note in the image above that the groom's name, "Jack Hancock" shows up twice in the marriage license in the blanks, once in the affidavit, and once in each signature line. The bride's name, "Katherine Brown" also shows up multiple times. System-builders were authorized to use the multiple appearances of information to increase the accuracy of their system hypotheses.

2.4 French Parish Collections

The French parish records (see Figure 1D) is a worst-case scenario for automatic HR transcription. These are free-form log books of French priests as they recorded the christenings, marriages, and burials of their parishioners. Similar to the Arkansas marriage records, only certain pieces of information on each page are genealogically relevant. However, what makes this collection particularly difficult is that most of the key genealogical facts appear in no particular location and with no particular format. There is also no repetition of facts and the information is in yet another language different from the other three collections.



N° 29
Baptême de
Désiré Pierre
Germain

[illegible]

Figure 1D. French Parish Records

3. PREPARING FOR SCORING

3.1. IRIS Metrics

Word error rate (WER) is a common metric used for the automatic transcription of media, so it or a variant seems an appropriate score for use in IRIS. WER divides the total number of substitution, deletion, and insertion word errors by the number of words in a perfect transcription. Substitutions and deletions are merely replacements or removal of words from the perfect transcript, but insertions are extra words that do not belong. Due to insertions, it is possible for a system to have WER in excess of 100%. This will become relevant later.

WER treats all words as equal. For genealogical purposes, though, all words do not have equal value. Gender words like “M” or “F,” though valuable, are not as genealogically beneficial as personal name pieces (like “Samuel” or “Schmidt”) or locative name pieces (such as “Boston” or “Colorado”). Consequently, IRIS chose to weight personal name pieces as five points each and locative name pieces as two points. All other words were treated as one point.

For censuses, there were also header components such as the location of the enumeration district. For end use, it would be critical for this information to be associated with each of the census rows, so accurate transcription of the header would be critical. These header pieces were therefore counted as if they had appeared on every row of the census, but the header place words were only counted as one point and all other header words were counted as 0.5 points.

3.2. Flexible Evaluation Systems

Since systems were required to automatically transcribe documents without being given any reference regions for what to transcribe on a given page, it would become very possible for a system to skip or partially transcribe rows and columns. One could easily imagine an HR system which failed to identify the first row of a US census image and then perfectly transcribed the remaining 49 rows. A human would likely treat the error for such as a situation as about 2% in that 49/50 rows were properly transcribed. Yet if the system expected that the 49 row transcripts corresponded to the *first* 49 rows, it might believe that there is 100% error or more. To account for this, IRIS created a flexible scorer which would attempt to maximize the score through minimum edit distance of the rows (i.e., try to align the rows so as to give participant sites the best score possible).

Even within a properly-identified cell, such as a given name field, the system may have believed there was a name “George E” when the actual transcript was only “George.” The IRIS scoring tool also tries to maximally align each field to give the most credit possible.

Dittos are a serious problem in census collections. A census-taker could record the last name of the father as “Jones” and then use hash marks (‘’) on subsequent lines to indicate that other family members also share the name “Jones.” Sometimes, merely a dash (---) or a blank space would be used to indicate this replication. In fact, in the US census, this is a particular problem because the data is represented as surnames followed by given names within the same field with no box line or delimiter, so a name like “ George Washington,” where the space was deliberately left as a ditto from the previous line might lead a HR system to believe that the name should be /George/ Washington. The scorer was thus constructed to account for these and other similar kinds of conditions.

Lastly, due to the inaccuracies and inconsistencies in the training and evaluation transcripts, as well as to the imposed transcription standards, IRIS sought to compensate for common variations that might appear in the “perfect” transcripts versus the HR system hypotheses. A painstaking review was conducted of all the possible results to ensure that name variations did not result in system penalties. The system was therefore built to accommodate single and multiword name variants which could easily be confusable by systems or humans (such as “Hernandez” vs. “Hernandes”); punctuation (“J.” vs “J”); segmentation issues (such as “Mc Donald” vs. “McDonald”); and differences between what appears on the actual page as opposed to what the transcription guidelines required (“PA” vs “Pennsylvania,” or “F” vs “Female”).

4. EVALUATION RESULTS

4.1. IRIS Participants

As was mentioned earlier, the call for this evaluation was extended to various research organizations. Some of the interested organizations later reported that they did not have HR systems that were sophisticated enough to tackle these challenges, and others received the data but were not able to generate any results. Two organizations, however, were able to produce results, namely A2iA and BBN.

A2iA is a Paris-based company which got its start in the early 1990s working on the automatic transcription of the handwritten parts of bank checks as well as postal addresses. By the start time of evaluation, A2iA had some preliminary experience working with the 1920 US Census and multilingual documents (see [3]).

BBN, a subsidiary to Raytheon, is a Boston-based company which has been a major US government contractor with special emphases in text and media processing. BBN was selected as one of the major contractors in the MADCAT effort mentioned earlier [4].

4.2. Best and Human Performance Per-Collection

For each of the IRIS collections, participants were allowed to submit as many variations of their systems as desired. The best result of those would be treated as the leading contribution from the participant site.

For this paper, we wish to present the best-performing result for each of the particular *collections*. Table 2 shows the best per-collection WWER scores across all of the system variations.

Collection	**Average Per-Record WWER**	**Minimum Per-Record WWER**	**Maximum Per-Record WWER**	**Std Dev Per-Record WWER**
1930 US Census	**19.6%**	2.73%	98.4%	12.8%
1930 Mexico Census	**47.4%**	5.59%	374%	30.9%
Arkansas Marriages	**29.4%**	0.00%	103%	18.4%
French Parish	**92.4%**	22.0%	198%	12.1%

Table 2: Best-performing System WWER per Collection

The primary goal of IRIS was to determine how close HR is to being able to transcribe genealogically-relevant documents. Since Table 2 shows the world's best-scoring results in the IRIS collections, it is relevant to know how close these scores are to human levels of performance.

For the IRIS transcripts that were provided for system development, the transcripts were a product of three separate volunteers. The first two volunteers, which we will refer to here as A and B, are given the task of transcribing the document independently. The third volunteer serves as an arbitrator (ARB) who is asked to vet the A and B transcripts and fix any errors.

We cannot readily evaluate human performance of the non-English data nor the Arkansas data because their "gold standards" were taken from the ARB transcript which in turn was derived from the A and B transcripts. Yet we can use the A-side transcript to measure performance on the US 1930 Census because systems on that collection were scored against the newly-created 99.5%-accurate transcripts. Table 3 shows the human levels of performance on the US Census. It also shows that for this collection, the system error rate is only about 2.48 times that of human error. This suggests automation is not far off from reaching human accuracies.

Collection	**Average Per-Record WWER (Human)**	**Automatic / Human Ratio of Average Per-Record WWER**
1930 US Census	**7.9%**	2.48

Table 3: Human WWER on US Collection

4.3. Per-Category Results of Best System Results

Though the overall performance is quite good, particularly for the US Census, it is clear that there would need to be some investment to move current error rates down to the level of human error rates. Even so, an analysis of individual fields on a per-collection basis might reveal that HR systems are ready for immediate production usage on some fields. For example, a system could be useful for transcribing a particular field if it either has high *accuracy* (1-WWER) or, if whenever it makes a hypothesis, it will typically predict correctly (i.e., it has high *precision*). Table 4 shows the accuracy and the precision for each field in the order of appearance in the 1930 US Census for the best-performing system. The fields that perform with greater than 90% precision are in boldfaced.

FIELD	**ACCURACY (1-WWER)**	**PRECISION**
census_district (H)	0.362	0.373
census_county (H)	0.649	0.741
sheet_number (H)	0.724	0.742
sheet_ltr (H)	**0.976**	**0.992**
household_id	0.747	0.825
pr_name_full	0.813	0.840
pr_relationship	**0.910**	**0.940**
pr_sex	**0.943**	**0.961**
pr_race_or_color	**0.946**	**0.969**
pr_age	0.840	0.857
marital_status	**0.939**	**0.957**
pr_birthplace	0.757	0.864
pr_fthr_birthplace	0.771	0.874
pr_mthr_birthplace	0.776	0.877

Table 4: Field performance for Best US Census System

This same trend follows through with the other collections. For the Mexico data, sex_code, relationship_to_head, principal person's age, and marital status had precisions above 90%. For the Arkansas data, event_type has 90%+ precision (as does "groom's given name," but this could be due to leveraging the repeat appearance of the groom's name, as described previously). In the French parish data, none of the fields had precisions in excess of 90%, but the best-performing field was sex_code.

From this table and these other results, we can see that the HR systems perform quite well in the fields where a small vocabulary is required. Perhaps automation could therefore be used for these particular fields and humans could tag the more complex fields, like places.

4.4. Major Error Types

Table 2 showed that whereas there were some documents where systems had perfect transcription accuracies, there were also documents where the system had about 100% error or more. These kinds of failures could have serious ramifications for utility, so it is profitable to look at the conditions that resulted in these serious errors.

In the US Census data, records from Vermont typically caused serious HR errors. Figure 2A indicates one of these documents. Even when this image shows up on a computer screen, the text is severely faded in comparison to the template text. With effort, a human can transcribe this faded image, but the systems seem to be less prepared for this lower signal-to-noise ratio. In fact, if this single state had been thrown out of the test set, the overall error rate would have been reduced by almost 1% absolute.

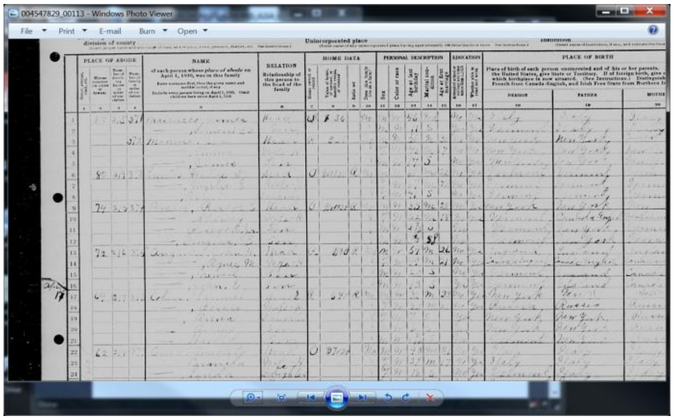
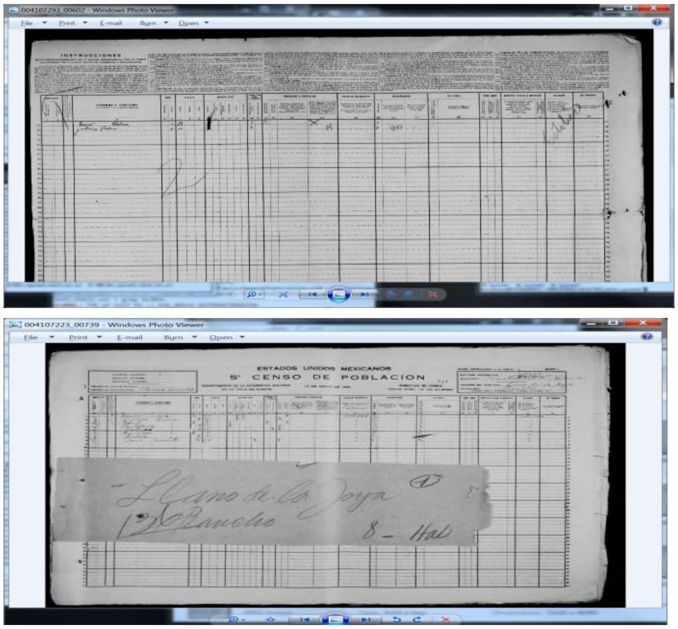


Figure 2A: Leading Error Type on US Census

When the systems were applied to the Mexico Census data, each participant's engine had major errors with different kinds of images. Figures 2B and 2C show images that produce faulty recognition results. For one of the systems, Figure 2B was transcribed as having data in every row – probably due to the handling of dittos. This led to a high insertion penalty and an error rate in excess of 300%. For the other system, images similar to Figure 2C were a problem due to unexpected masking on the page. Understandably, in this latter situation, the system produced transcription components that were not in the actual census page. Since these were infrequent issues, they had little effect on the overall average WWER score.



Figures 2B and 2C: Leading Error Types on Mexico Data

What about error on the held-out states? It was mentioned previously that there was a US and Mexican state that were held out of the training data but were represented in the evaluation data. The additional states had little negative additional contribution to the error.

4.5. Minor Error Types

The core, repetitive system issues which contributed to much of the overall error was due to several factors. Systems tended to err on the side of not hypothesizing anything when in doubt, so this led to many deletions. Likewise, there were common errors like: Missouri vs. Mississippi; Carolina vs. South Carolina; and China vs. Chihuahua. Although these deletions and substitutions are definite errors, it is evident that systems are not far wrong.

5. CONCLUSIONS

This evaluation demonstrates that handwriting recognition systems are extremely close to providing full-document-transcription value for genealogical records in English and are likely only a few research-years away from providing this kind of value in non-English languages. Even today, the systems could be profitably applied to the transcription of fields with limited vocabularies.

REFERENCES

[1] http://www.darpa.mil/Our_Work/I2O/Programs/Multi-lingual_Automatic_Document_Classification,_Analysis_ and_Translation_%28MADCAT%29.aspx

[2] http://www.itl.nist.gov/iad/mig/tests/hart/OPENHART_Reports_20101015-1436/

[3] http://www.a2ia.com//Web_Bao/History-Eng.aspx

[4] http://bbn.com/news_and_events/press_releases/2010/pr_madcat_070710