An Unsupervised Model of Orthographic Variation for Historical Document Transcription

Dan Garrette
Computer Science & Engineering
University of Washington
dhg@cs.washington.edu

Hannah Alpert-Abrams
Comparative Literature Program
University of Texas at Austin
halperta@gmail.com

Abstract

Historical documents frequently exhibit extensive orthographic variation, including archaic spellings and obsolete shorthand. OCR tools typically seek to produce so-called *diplomatic* transcriptions that preserve these variants, but many end tasks require transcriptions with *normalized* orthography. In this paper, we present a novel joint transcription model that learns, unsupervised, a probabilistic mapping between modern orthography and that used in the document. Our system thus produces dual diplomatic and normalized transcriptions simultaneously, and achieves a 35% relative error reduction over a state-of-the-art OCR model on diplomatic transcription, and a 46% reduction on normalized transcription.

1 Introduction

Optical Character Recognition (OCR) for historical texts, a challenging problem due to unknown fonts and deteriorating documents, is made even more difficult by the fact that orthographic conventions including spelling, accent usage, and shorthands have not been consistent across the history of printing. For this reason, modern language models (LMs) yield poor performance when trying to recognize characters on the pages of these documents. Furthermore, transcription of the actual printed characters may not always be the most desirable output.

Greg (1950) describes two types of transcription: one that preserves variants and typographical errors, and another that records the *substantive* content, with this noise removed. Though in 1950 the substantive version was the norm, today these have become two distinct but equally valid tasks. *Diplomatic* transcription, the standard in contemporary OCR, preserves the variants of the document valuable to book historians and linguists. *Normalized* or modernized transcription recovers the substantive content, producing a text that adheres to modern standards. Normalized transcriptions are easier for users to read, and make large collections of historical texts indexable and searchable (Driscoll, 2006).

The current ideal for digital editions of historical texts has been described as a combination of diplomatic and normalized transcription (Pierazzo, 2014). This is generally achieved with a pipeline: first OCR is used to transcribe the document, then an (often manual) post-hoc normalization is performed. However, such a pipeline will result in cascading errors from OCR mistakes, and fails to make use of knowledge about modern language during the initial transcription. Additionally, post-processing tools are typically cumbersome language-specific, hand-built systems (Baron and Rayson, 2008; Burns, 2013; Hendrickx and Marquilhas, 2011).

In this work, we introduce a novel OCR model designed to jointly produce both diplomatic and normalized transcriptions. The model is an extension of Berg-Kirkpatrick et al.'s (2013) *Ocular*, the state of the art in historical OCR. Ocular's innovative ability to handle the material challenges of OCR (unknown fonts, uneven inking, etc.) depends on its use of a character n-gram LM. Our model improves the quality of Ocular's transcriptions by automatically learning a probabilistic mapping between the LM, which is trained on modern text, and the unique orthography of the document. This results in both an

Proceedings of NAACL-HLT 2016, pages 467–472,
San Diego, California, June 12-17, 2016.

improved orthographically-correct diplomatic transcription and a modern-style normalized transcription. To our knowledge, this represents the first OCR system that jointly produces both diplomatic and normalized transcriptions.

We evaluate our model on a multilingual collection of books exemplifying a high degree of orthographic variation. For diplomatic transcription, our unsupervised joint model achieves an error reduction of 35% over the baseline Ocular system without support for orthographic variation, and nearly matches the error rate of an approach proposed by earlier work that uses a hand-constructed ruleset of orthographic rewrites. However, for the new task of normalized transcription, we achieve a 46% error reduction over the baseline, as well as a 28% reduction over the hand-built ruleset approach.

2 Data

The *Primeros Libros* corpus dataset introduced in our previous work consists of multilingual (Spanish/Latin/Nahuatl) books printed in Mexico in the 1500s (Garrette et al., 2015). The original dataset includes gold diplomatic transcriptions of pages from five books of different time periods, regions, languages, and fonts. We additionally include two new monolingual Spanish *Primeros Libros* books annotated with both diplomatic and normalized transcriptions. Spanish-only texts were needed in order to find language-competent annotators skilled enough to create the more challenging normalized transcriptions. We used each of the seven books in isolation since they each have a different font. For each book, we used 20 pages for training and 10 for testing. For two of the books, an additional 10 pages were held out for tuning hyperparameters with grid search.

To produce the Spanish and Latin LMs, we used texts from Project Gutenberg; these documents were written during the target historical period, but all follow modernization standards including substitution for obsolete characters and expansion of shorthand. These texts were chosen because they are a realistic sample set that is freely and publicly available. In the Nahuatl case, scarce online resources made it necessary to supplement Project Gutenberg text with that from a private collection.

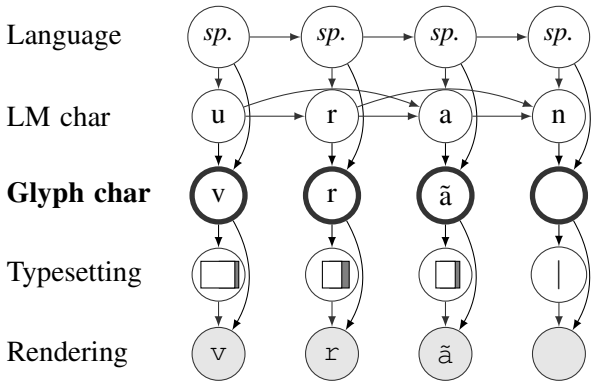


Figure 1: Our generative OCR model with the new *glyph* layer (bolded). The Spanish (*sp.*) n-gram language model (LM) generates a sequence of characters according to standard Spanish spellings, *uran* in this case, from the word *proc**uran**do* which may be written *proc**vrã**do*. Language-specific character-replacement probabilities are used to generate a *glyph char* from each LM char, producing *vrã* and a zero-width (elided) *n*. Finally, the model generates a bounding box and right-side padding (the typesetting) and a pixel-rendering of the glyph character.

3 Model

We extend Ocular, the generative model of Berg-Kirkpatrick et al. (2013), and its EM training procedure, to support our unsupervised approach to jointly modeling both diplomatic and normalized transcription. Ocular works by modeling the operation of a hand press in order to learn unknown fonts in the presence of the visual noise of the printing process: uneven inking and spacing in particular. Briefly, Ocular's generative story is as follows. First, a sequence of language states ℓ_i is generated according to $P^{\text{LANG}}(\ell_i \mid \ell_{i-1})$, where ℓ_{i-1} and ℓ_i may only differ on the start of a word. For each state, a character c_i is generated according to its language-specific n-gram LM: $P^{\text{CHAR}}_{\ell_i}(c_i \mid c_{i-n+1} \ldots c_{i-1})$. Next, a state's *typesetting* t_i, consisting of a character's bounding box, vertical offset, and between-character padding, is generated according to $P^{\text{TYPE}}(t_i \mid c_i)$. Finally, the character image is rendered as a pixel-matrix inside the bounding box: $P^{\text{REND}}(x_i \mid c_i, t_i)$.[1]

A major downside to the Ocular model is that ren-

[1] See previous work for fuller detail including how a typesetting is composed of its parts, and how pixels are generated.

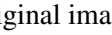
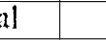
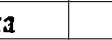
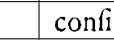
	char sub. (c → q)	char sub. (s → long s)	elision (que → q̃)	accent drop (ó → o)	doubled (c → cc)	typo (e → r)
Original image						
Baseline trans.	**q**ual	eña	á ól	confideraci**o**n	pe**cc**ados	Primeram**i**nte
Our diplomatic trans.	**q**ual	e**ſ**ta	a**q̃**l	confideraci**o**n	pe**cc**ados	Primeram**r**nte
Our normalized trans.	**c**ual	es**t**a	a**que**l	consideraci**ó**n	pe**c**ados	Primeram**e**nte

Table 1: Examples of automatic diplomatic and normalized transcriptions taken from actual system output.

dered image x_i is always generated directly from LM character c_i, resulting in transcription errors when printed characters don't follow the spellings in the LM. Our model (Figure 1) adds an additional layer to the generative model that de-couples the LM from the rendering by allowing the LM-generated character c_i to be replaced by a possibly different *glyph character* g_i which is rendered instead.

For the generative story of our new model, we again begin by generating pairs (ℓ_i, c_i). However, instead of typesetting c_i, we generate a distinct glyph character g_i as its replacement, according to $P^{\text{GLYPH}}_{\ell_i}(g_i \mid c_i)$. Orthographic substitution patterns are language-specific, and thus P^{GLYPH} is as well. Finally, we typeset and render g_i (instead of c_i) using $P^{\text{TYPE}}(t_i \mid g_i)$ and $P^{\text{REND}}(x_i \mid g_i, t_i)$.

We follow the previous Ocular work for the definitions of P^{LANG}, $P^{\text{CHAR}}_{\ell_i}$, P^{TYPE}, and P^{REND}. We define the new conditional distribution $P^{\text{GLYPH}}_{\ell_i}$, specifying the probability of rendering g when the LM generated c, given that the language is ℓ, as follows:

$$P^{\text{GLYPH}}_{\ell}(g \mid c) = \begin{cases} (1-\kappa) + \kappa \cdot p(g \mid c, \ell) & \text{if } g = c \\ \kappa \cdot p(g \mid c, \ell) & \text{else} \end{cases}$$

Constant κ defines a Bernoulli parameter specifying the fixed probability of deterministically choosing to render c_i directly (i.e., $g_i = c_i$). We set $\kappa = 0.9$ to bias the model away from substitutions in general. The remaining $(1 - \kappa)$ probability mass is then divided among all potential output glyph characters.

Table 1 shows some of the common substitution patterns that our model addresses. For a direct rendering of c_i, a letter substitution, or the dropping of an accent, g_i will be a simple character drawn from the language's set of valid characters (each language may have a different set of permitted characters, e.g. accented letters). In order to support the tilde-elision shorthand, we permit g_i to be a tilde-annotated version of c_i, and for doubled letters, we permit g_i to be $c_i c_i$, for which we typeset and render c_i twice. Finally, to allow for elided letters, including the dropping of a line-break hyphen, we allow g_i to be a special ELISION glyph that renders only as a zero-width space character.

The parameters of the glyph substitution model are learned in an unsupervised fashion as part of Ocular's EM procedure via a hard parameter update:

$$p(g \mid c, \ell) = \frac{\mathit{freq}(\ell, c, g) + 1}{\sum_{g'} (\mathit{freq}(\ell, c, g') + 1)}$$

where $\mathit{freq}(\ell, c, g)$ is the number of times in the training iteration that the model chose to replace c with g in a word (automatically) determined to be of language ℓ. The $+1$ term is Laplace smoothing.

To guide the model and improve efficiency, we employ a number of constraints governing which kinds of substitutions are valid. Among these, we stipulate that substitutions must be letter-to-letter, diacritics may only be added to lowercase letters, only *s* can replace long-*s*, and elision-tilde-marked letters must be followed by one or more elisions.

4 Experiments

As a first baseline, we compare against Ocular with no orthographic variation handling, in which characters generated by the LM are rendered directly.

As a second baseline, we compare to our previous work, which improved Ocular's diplomatic transcription accuracy by introducing orthographic variation directly into the LM with hand-constructed language-specific orthographic rules to rewrite the LM training data prior to n-gram estimation (Garrette et al., 2015). However, this rule-based preprocessing approach is inadequate in many ways. First, annotators do not know the full range of orthographic variations, or their frequencies, in each document, and it is impossible to write rules to handle typos. Furthermore, a highly language-proficient

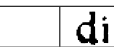
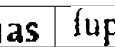
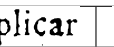
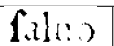
Original image						
Baseline trans.	dias	fuplicar	faltro	alaba	tí pues	Nefuxpo
Our diplomatic trans.	*días	ſuplicar	*faliio	alabã	ti pues	*Jeſuxpo
Our normalized trans.	días	*súplicar	*salió	*alabar	*te pues	*Jesuxpo
Gold diplomatic trans.	dias	ſuplicar	ſaluo	alabã	ti pues	Jeſu xp̃o
Gold normalized trans.	días	suplicar	salvo	alaban	ti pues	Jesu Cristo

Table 2: Actual system outputs containing transcription errors. Our incorrect outputs are starred (*).

Orthographic variation strategy	Diplomatic CER	Diplomatic WER	Normalized CER	Normalized WER
No handling	13.2	45.7	17.4	47.6
Hand-written rules	**8.5**	**30.8**	13.1	37.9
Unsupv. joint model	8.6	32.7	**9.5**	**27.6**

Table 3: Experimental results for both Diplomatic (preserving variation) and Normalized (modern orthography) transcription tasks. Results given as both character error rate (CER, including punctuation) and word error rate (WER, without punctuation).

c	g	$freq(sp., c, g)$	$P^{\text{GLYPH}}_{spanish}(g \mid c)$
-	ELIDED	52	0.0881
ó	o	31	0.0526
s	ſ (long s)	325	0.0352
q	q̃	9	0.0222
n	ELIDED	57	0.0136
v	u	55	0.0129
o	õ	20	0.0091
c	cc	23	0.0028

Table 4: A sample of high-probability Spanish substitution rules learned by our unsupervised model.

annotator is required, which is not always feasible with, e.g, rare indigenous languages.

Each baseline model has a single output, evaluated against both diplomatic and normalized gold.

Our source code and evaluation data are freely available at *https://github.com/tberg12/ocular* and *https://github.com/dhgarrette/ocr-evaluation-data*.

5 Results and Discussion

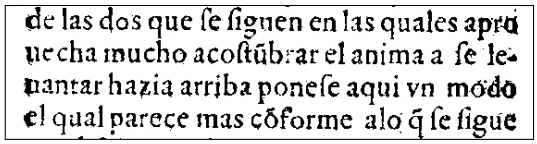
Our results are shown in Table 3 and some correct example system outputs can be seen in Table 1.

Our model's accuracy in producing diplomatic transcriptions is substantially better than baseline Ocular performance, yielding a 35% relative character error rate reduction. Further, our model's diplomatic transcription accuracy is comparable to the LM replacement-rule approach of our previous work, but achieves this result with only unsupervised learning, as opposed to expert-produced rules. We can see in Table 4 that our model is able to discover and apply appropriate probabilities to relevant substitution rules. For the new normalized-transcription task, even larger gains were achieved: a 46% relative error reduction over Ocular, and a 28% reduction over the rule-based approach.

5.1 Error Analysis

Table 2 displays a sample of errors in our system output. (1) The word *días* is printed without an accent though it has one in modern Spanish. Our system is unable to distinguish between a dot and an accent above the *i*, and thus it opts to output the accented version since it is preferred by the LM. (2) The word *suplicar* does not have any accents in modern Spanish, but the model is over-eager in this case and attempts to revive an accent where there should not be one. (3) The word *salvo* is printed here as the variant *saluo*, but its under-inking leaves the *u* in disconnected pieces, resembling a pair of *i* characters. The LM model believes this to be the (valid) word *salió* with the accent dropped and the *i* doubled. (4) The model guesses incorrectly that the elision at the end of *alabã* is an *r* since *alabar* is a valid word. (5) The model correctly recognizes that the letters *ti* are printed, but the LM believes the normalized form is *te* even though *te pues* is not valid Spanish, perhaps because *te puedo* is a very common phrase and the six-gram context isn't enough to make that distinction. (6) Finally, there are some special idiomatic shorthands that our model is simply unable to understand because they have no clear

Original Image



Diplomatic

de las dos que **ſ**e **ſ**iguen en las quales apro
uecha mucho aco**ſ**t**ũ**brar el **a**nima á **ſ**e le-
pantar ha**z**ia arriba p**o**ne**ſ**e aqu**i** **v**n modo
el **q**ual parece más c**õ**forme á lo **q̃** **ſ**e **ſ**igue

Normalized

de las dos que **s**e **s**iguen en las quales
apro**v**echa mucho aco**st****um**b**r**ar el **á**nima á **s**e
le**v**antar ha**c**ia arriba p**ó**ne**s**e aqu**í** **u**n modo
el **c**ual parece más c**on**forme á lo **que** **s**e **s**igue

Table 5: A document excerpt along with actual system outputs for both diplomatic and normalized transcriptions. Note that the normalization recovers the first line's missing line-break hyphen, allowing the full word *aprovecha* to be reassembled.

connection to what they are replacing. Here, *xp̃o* (from Greek letters *chi rho*) is shorthand for *Cristo*.

6 Conclusion

In this paper we presented a novel unsupervised OCR model for the joint production of diplomatic (variant-preserving) and normalized transcriptions of historical documents. The model is able to automatically learn a probabilistic mapping from a LM trained on modern text to the orthographic variants present in the document. This has the dual result of both considerably improving diplomatic transcription accuracy, while also enabling the model to simultaneously produce a normalized transcription.

Our model has the potential to significantly impact the work of scholars and librarians who wish to make digital texts easier to read, index, search, and study. Our approach also has the fortunate side-effect of producing metadata about orthographic variation in printed documents that may be valuable to scholars of book history. With our model, we are able to automatically induce sets of variation patterns used by printers, and the locations in the texts where those variants appear, without the need for labor-intensive page-by-page reading. Further, these induced mappings have probabilities attached, and are not simple rulebanks like those used in existing normalization work (Garrette et al., 2015; Baron and Rayson, 2008). Table 4, above, shows a sample of rules and their frequencies that resulted from training our model on Spanish documents.

Finally, our work helps to bridge the gap between historical text and mainline NLP. Orthographic variation lowers the accuracy of NLP tools due to high out-of-vocabulary rates and mismatched morphological features (Piotrowski, 2012). This is especially true when these tools are trained on the modern texts of the standard corpora used in NLP (Yang and Eisenstein, 2016). Normalization of historical texts have been shown to improve the quality of, for example, taggers (Rayson et al., 2007; Yang and Eisenstein, 2016) and parsers (Rayson et al., 2007). These techniques mirror those applied to the processing of text in social media, such as Twitter, where there is a high degree of slang and shorthand (Gimpel et al., 2011; Eisenstein, 2013; Yang and Eisenstein, 2013). Most approaches train off-the-shelf NLP tools on modern text and then apply normalization techniques to historical texts to transform them into something resembling the modern training input (Scheible et al., 2011). A joint model such as ours that automatically learns orthographic variations while training the NLP model might overcome some of the limitations of using such a pipeline approach.

Acknowledgments

We would like to thank Stephanie Wood, Kelly McDonough, Albert Palacios, Adam Coon, Sergio Romero, and Kent Norsworthy for their input, advice, and assistance on this project. We would also like to thank Taylor Berg-Kirkpatrick, Dan Klein, and Luke Zettlemoyer for their valuable feedback on earlier drafts of this paper. This work is supported in part by a Digital Humanities Implementation Grant from the National Endowment for the Humanities for the project Reading the First Books: Multilingual, Early-Modern OCR for Primeros Libros.

References

Alistair Baron and Paul Rayson. 2008. VARD2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of The Postgraduate Conference in Corpus Linguistics*.

Taylor Berg-Kirkpatrick and Dan Klein. 2014. Improved typesetting models for historical OCR. In *ACL*.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of ACL*.

Philip R. Burns. 2013. MorphAdorner v2: A Java library for the morphological adornment of English language texts. `https://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf`.

Arthur P. Dempster, Nan M. Laird, and Donald. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39.

Thomas G. Dolan. 2012. The Primeros Libros Project. *The Hispanic Outlook in Higher Education*, 22:20–22, March.

M. J. Driscoll. 2006. Electronic textual editing: Levels of transcription. In Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth, editors, *Electronic Textual Editing*. Modern Language Association.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL*.

Dan Garrette, Hannah Alpert-Abrams, Taylor Berg-Kirkpatrick, and Dan Klein. 2015. Unsupervised code-switching for multilingual historical document transcription. In *Proceedings of NAACL*.

Kevin Gimpel, Nathan Schneider, Brendan OConnor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, , and Noah A. Smith. 2011. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of ACL*.

W. W. Greg. 1950. The rationale of copy-text. *Studies in Bibliography*, 3:19–36.

Iris Hendrickx and Rita Marquilhas. 2011. From old texts to modern spellings: An experiment in automatic normalisation. *Journal of Language Technology and Computational Linguistics*, 26(2):65–76.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*.

Elena Pierazzo. 2014. Digital documentary editions and the Others. *Scholarly Editing*, 35:1–23.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5:1–157.

Primeros Libros. 2010. Los Primeros Libros de las Américas: Impresos Americanos del siglo XVI en las bibliotecas del mundo. *http://primeroslibros.org/*.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the bard: Evaluating the accuracy of a modern POS tagger on early modern English corpora. In *Corpus Linguistics Conference*.

Silke Scheible, Richard J Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS-tagger on early modern German text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Yi Yang and Jacob Eisenstein. 2013. A log-linear model for unsupervised text normalization. In *Proceedings of EMNLP*.

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical English. In *Proceedings of NAACL*.