Improved Typesetting Models for Historical OCR

Taylor Berg-Kirkpatrick **Dan Klein**
Computer Science Division
University of California, Berkeley
{tberg,klein}@cs.berkeley.edu

Abstract

We present richer typesetting models that extend the unsupervised historical document recognition system of Berg-Kirkpatrick et al. (2013). The first model breaks the independence assumption between vertical offsets of neighboring glyphs and, in experiments, substantially decreases transcription error rates. The second model simultaneously learns multiple font styles and, as a result, is able to accurately track italic and non-italic portions of documents. Richer models complicate inference so we present a new, streamlined procedure that is over 25x faster than the method used by Berg-Kirkpatrick et al. (2013). Our final system achieves a relative word error reduction of 22% compared to state-of-the-art results on a dataset of historical newspapers.

1 Introduction

Modern OCR systems perform poorly on historical documents from the printing-press era, often yielding error rates that are too high for downstream research projects (Arlitsch and Herbert, 2004; Shoemaker, 2005; Holley, 2010). The two primary reasons that historical documents present difficultly for automatic systems are (1) the typesetting process used to produce such documents was extremely noisy and (2) the fonts used in the documents are unknown. Berg-Kirkpatrick et al. (2013) proposed a system for historical OCR that generatively models the noisy typesetting process of printing-press era documents and learns the font for each input document in an unsupervised fashion. Their system achieves state-of-the-art results on the task of historical document recognition.

We take the system of Berg-Kirkpatrick et al. (2013) as a starting point and consider extensions of the typesetting model that address two shortcomings of their model: (1) their layout model assumes that baseline offset noise is independent for each glyph and (2) their font model assumes a single font is used in every document. Both of these assumptions are untrue in many historical datasets.

The baseline of the text in printing-press era documents is not rigid as in modern documents but rather drifts up and down noisily (see Figure 2). In practice, the vertical offsets of character glyphs change gradually along a line. This means the vertical offsets of neighboring glyphs are correlated, a relationship that is not captured by the original model. In our first extension, we let the vertical offsets of character glyphs be generated from a Markov chain, penalizing large changes in offset. We find that this extension decreases transcription error rates. Our system achieves a relative word error reduction of 22% compared to the state-of-the-art original model on a test set of historical newspapers (see Section 4.1), and a 11% relative reduction on a test set of historical court proceedings.

Multiple font styles are also frequently used in printing-press era documents; the most common scenario is for a basic font style to co-occur with an italic variant. For example, it is common for proper nouns and quotations to be italicized in the Old Bailey corpus (Shoemaker, 2005). In our second extension, we incorporate a Markov chain over font styles, extending the original model so that it is capable of simultaneously learning italic and non-italic fonts within a single document. In experiments, this model is able to detect which words are italicized with 93% precision at 74% recall in a test set of historical court proceedings (see Section 4.2).

These richer models that we propose do increase the state space and therefore make inference more costly. To remedy this, we streamline inference by replacing the coarse-to-fine inference scheme of Berg-Kirkpatrick et al. (2013)

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 118–123,
Baltimore, Maryland, USA, June 23-25 2014.

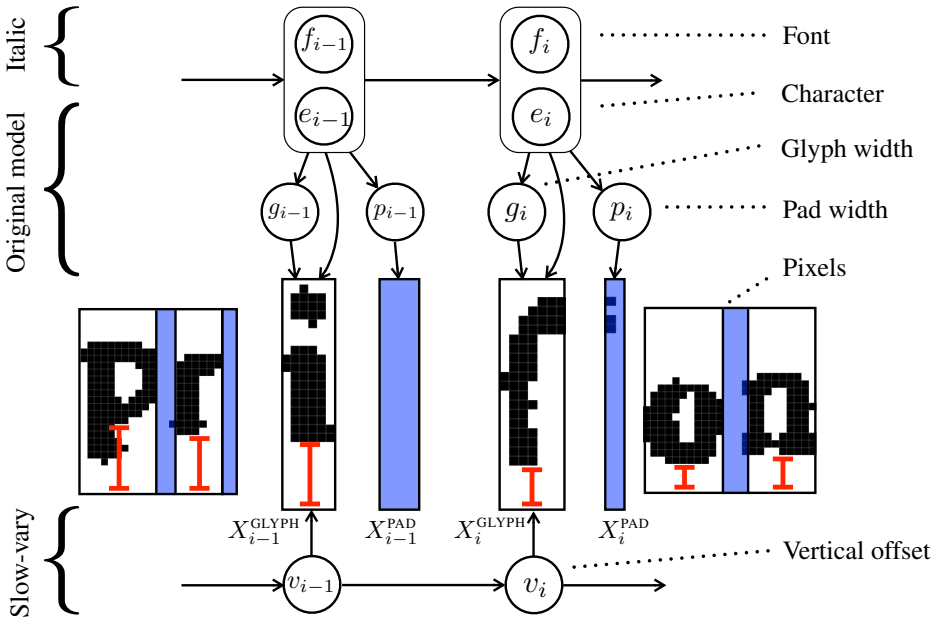

Figure 1: See Section 2 for a description of the generative process. We consider an extension of Berg-Kirkpatrick et al. (2013) that generates v_i conditioned on the previous vertical offset v_{i-1} (labeled Slow-vary) and an extension that generates a sequence of font styles f_i (labeled Italic).

with a forward-cost-augmented beaming scheme. Our method is over 25x faster on a typical document, yet actually yields *improved* transcriptions.

2 Model

We first describe the generative model used by the 'Ocular' historical OCR system of Berg-Kirkpatrick et al. (2013)[1] and then describe our extensions. The graphical model corresponding to their basic generative process for a single line of text is diagrammed in Figure 1. A Kneser-Ney (Kneser and Ney, 1995) character 6-gram language model generates a sequence of characters $E = (e_1, e_2, \ldots, e_n)$. For each character index i, a glyph box width g_i and a pad box width p_i are generated, conditioned on the character e_i. g_i specifies the width of the bounding box that will eventually house the pixels of the glyph for character e_i. p_i specifies the width of a padding box which contains the horizontal space before the next character begins. Next, a vertical offset v_i is generated for the glyph corresponding to character e_i. v_i allows the model to capture variance in the baseline of the text in the document. We will later let v_i depend on v_{i-1}, as depicted in Figure 1, but in the baseline system they are independent. Finally, the pixels in the ith glyph bounding box X_i^{GLYPH} are generated conditioned on the character e_i, width g_i, and vertical offset v_i, and the pixels in the ith pad bounding box X_i^{PAD} are generated conditioned on the width p_i. We refer the reader to Berg-Kirkpatrick et al. (2013) for the details of the pixel generation process. We have omitted the token-level inking random variables for the purpose of brevity. These can be treated as part of the pixel generation process.

Let X denote the matrix of pixels for the entire line, $V = (v_1, \ldots, v_n)$, $P = (p_1, \ldots, p_n)$, and $G = (g_1, \ldots, g_n)$. The joint distribution is written:

$$
\begin{aligned}
P(X,&V,P,G,E) = \\
&P(E) && \text{[Language model]} \\
&\cdot \prod_{i=1}^{n} P(g_i|e_i;\Phi) && \text{[Glyph widths]} \\
&\cdot \prod_{i=1}^{n} P(p_i|e_i;\Phi) && \text{[Pad widths]} \\
&\cdot \prod_{i=1}^{n} P(v_i) && \text{[Vertical offsets]} \\
&\cdot \prod_{i=1}^{n} P(X_i^{\text{PAD}}|p_i) && \text{[Pad pixels]} \\
&\cdot \prod_{i=1}^{n} P(X_i^{\text{GLYPH}}|v_i,g_i,e_i;\Phi) && \text{[Glyph pixels]}
\end{aligned}
$$

[1]The model we describe and extend has two minor differences from the one described by Berg-Kirkpatrick et al. (2013). While Berg-Kirkpatrick et al. (2013) generate two pad boxes for each character token, one to the left and one to the right, we only generate one pad box, always to the right. Additionally, Berg-Kirkpatrick et al. (2013) do not carry over the language model context between lines, while we do.

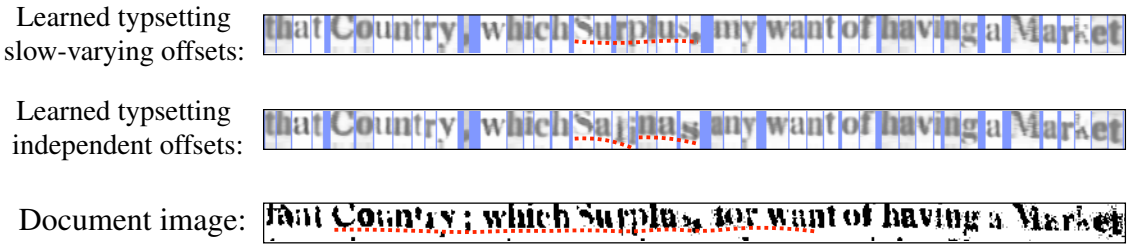

Figure 2: The first line depicts the Viterbi typesetting layout predicted by the OCULAR-BEAM-SV model. The second line depicts the same, but for the OCULAR-BEAM model. Pad boxes are shown in blue. Glyphs boxes are shown in white and display the Bernoulli template probabilities used to generate the observed pixels. The third line shows the corresponding portion of the input image.

The font is parameterized by the vector Φ which governs the shapes of glyphs and the distributions over box widths. Φ is learned in an unsupervised fashion. Document recognition is accomplished via Viterbi decoding over the character random variables e_i.

2.1 Slow-varying Offsets

The original model generates the vertical offsets v_i independently, and therefore cannot model how neighboring offsets are correlated. This correlation is actually strong in printing-press era documents. The baseline of the text wanders in the input image for two reasons: (1) the physical groove along which character templates were set was uneven and (2) the original document was imaged in a way that produced distortion. Both these underlying causes are likely to yield baselines that wander slowly up and down across a document. We refer to this behavior of vertical offsets as slow-varying, and extend the model to capture it.

In our first extension, we augment the model by incorporating a Markov chain over the vertical offset random variables v_i, as depicted in Figure 1. Specifically, v_i is generated from a discretized Gaussian centered at v_{i-1}:

$$P(v_i|v_{i-1}) \propto \exp\left(\frac{(v_i - v_{i-1})^2}{2\sigma^2}\right)$$

This means that the if v_i differs substantially from v_{i-1}, a large penalty is incurred. As a result, the model should prefer sequences of v_i that vary slowly. In experiments, we set $\sigma^2 = 0.05$.

2.2 Italic Font Styles

Many of the documents in the Old Bailey corpus contain both italic and non-italic font styles (Shoemaker, 2005). The way that italic fonts are used depends on the year the document was printed, but generally italics are reserved for proper nouns, quotations, and sentences that have a special role (e.g. the final judgment made in a court case). The switch between font styles almost always occurs at space characters.

Our second extension of the typesetting model deals with both italic and non-italic font styles. We augment the model with a Markov chain over font styles f_i, as depicted in Figure 1. Each font style token f_i takes on a value in $\{\text{ITALIC}, \text{NON-ITALIC}\}$ and is generated conditioned on the previous font style f_{i-1} and the current character token e_i. Specifically, after generating a character token that is not a space, the language model deterministically generates the last font used. If the language model generates a space character token, the decision of whether to switch font styles is drawn from a Bernoulli distribution. This ensures that the font style only changes at space characters.

The font parameters Φ are extended to contain entries for the italic versions of all characters. This means the shapes and widths of italic glyphs can be learned separately from non-italic ones. Like Berg-Kirkpatrick et al. (2013), we initialize the font parameters from mixtures of modern fonts, using mixtures of modern italic font styles for italic characters.

3 Streamlined Inference

Inference in our extended typesetting models is costly because the state space is large; we propose an new inference procedure that is fast and simple.

Berg-Kirkpatrick et al. (2013) used EM to learn the font parameters Φ, and therefore required expected sufficient statistics (indicators on (e_i, g_i, v_i) tuples), which they computed using coarse-to-fine inference (Petrov et al., 2008; Zhang and Gildea, 2008) with a semi-Markov dynamic program (Levinson, 1986). This approach is effec-

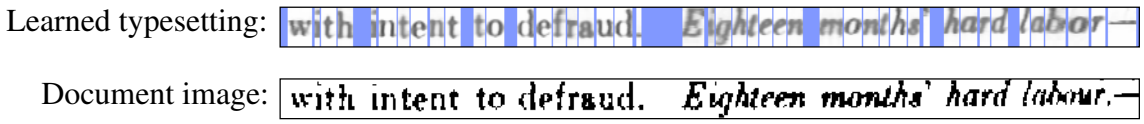

Figure 3: This first line depicts the Viterbi typesetting layout predicted by the OCULAR-BEAM-IT model. Pad boxes are shown in blue. Glyphs boxes are shown in white and display the Bernoulli template probabilities used to generate the observed pixels. The second line shows the corresponding portion of the input image.

tive, but slow. For example, while transcribing a typical document consisting of 30 lines of text, their system spends 63 minutes computing expected sufficient statistics and decoding when run on a 4.5GHz 4-core CPU.

We instead use hard counts of the sufficient statistics for learning (i.e. perform hard-EM). As a result, we are free to use inference procedures that are specialized for Viterbi computation. Specifically, we use beam-search with estimated forward costs. Because the model is semi-Markov, our beam-search procedure is very similar the one used by Pharaoh (Koehn, 2004) for phrase-based machine translation, only without a distortion model. We use a beam of size 20, and estimate forward costs using a character bigram language model. On the machine mentioned above, transcribing the same document, our simplified system that uses hard-EM and beam-search spends only 2.4 minutes computing sufficient statistics and decoding. This represents a 26x speedup.

4 Results

We ran experiments with four different systems. The first is our baseline, the system presented by Berg-Kirkpatrick et al. (2013), which we refer to as OCULAR. The second system uses the original model, but uses beam-search for inference. We refer to this system as OCULAR-BEAM. The final two systems use beam-search for inference, but use extended models: OCULAR-BEAM-SV uses the slow-varying vertical offset extension described in Section 2.1 and OCULAR-BEAM-IT uses the italic font extension described in Section 2.2.

We evaluate on two different test sets of historical documents. The first test set is called Trove, and is used by Berg-Kirkpatrick et al. (2013) for evaluation. Trove consists of 10 documents that were printed between 1803 and 1954, each consisting of 30 lines, all taken from a collection of historical Australian newspapers hosted by the National Library of Australia (Holley, 2010). The second test set, called Old Bailey, consists of 20 documents that were printed between 1716 and 1906, each consisting of 30 lines, all taken from a the proceedings of the Old Bailey Courthouse in London (Shoemaker, 2005).[2] Following Berg-Kirkpatrick et al. (2013), we train the language model using 36 millions words from the New York Times portion of the Gigaword corpus (Graff et al., 2007).[3]

4.1 Document Recognition Performance

We evaluate predicted transcriptions using both character error rate (CER) and word error rate (WER). CER is the edit distance between the guessed transcription and the gold transcription, divided by the number of characters in the gold transcription. WER is computed in the same way, but words are treated as tokens instead of characters.

First we compare the baseline, OCULAR, to our system with simplified inference, OCULAR-BEAM. To our surprise, we found that OCULAR-BEAM produced better transcriptions than OCULAR. On Trove, OCULAR achieved a WER of 33.0 while OCULAR-BEAM achieved a WER of 30.7. On Old Bailey, OCULAR achieved a WER of 30.8 while OCULAR-BEAM achieved a WER of 28.8. These results are shown in Table 1, where we also report the performance of Google Tesseract (Smith, 2007) and ABBYY FineReader, a state-of-the-art commercial system, on the Trove test set (taken from Berg-Kirkpatrick et al. (2013)).

Next, we evaluate our slow-varying vertical offset model. OCULAR-BEAM-SV out-performs OCULAR-BEAM on both test sets. On Trove, OCULAR-BEAM-SV achieved a WER of 25.6, and on Old Bailey, OCULAR-BEAM-SV achieved a WER of 27.5. Overall, compared to our baseline

[2] Old Bailey is comparable to the the second test set used by Berg-Kirkpatrick et al. (2013) since it is derived from the same collection and covers a similar time span, but it consists of different documents.

[3] This means the language model is out-of-domain on both test sets. Berg-Kirkpatrick et al. (2013) also consider a perfectly in-domain language model, though this setting is somewhat unrealistic.

system, OCULAR-BEAM-SV achieved a relative reduction in WER of 22% on Trove and 11% on Old Bailey.

By looking at the predicted typesetting layouts we can make a qualitative comparison between the vertical offsets predicted by OCULAR-BEAM and OCULAR-BEAM-SV. Figure 2 shows representations of the Viterbi estimates of the typesetting random variables predicted by the models on a portion of an example document. The first line is the typesetting layout predicted by OCULAR-BEAM-SV and the second line is same, but for OCULAR-BEAM. The locations of padding boxes are depicted in blue. The white glyph bounding boxes reveal the values of the Bernoulli template probabilities used to generate the observed pixels. The Bernoulli templates are produced from type-level font parameters, but are modulated by token-level widths g_i and vertical offsets v_i (and inking random variables, whose description we have omitted for brevity). The predicted vertical offsets are visible in the shifted baselines of the template probabilities. The third line shows the corresponding portion of the input image. In this example, the text baseline predicted by OCULAR-BEAM-SV is contiguous, while the one predicted by OCULAR-BEAM is not. Given how OCULAR-BEAM-SV was designed, this meets our expectations. The text baseline predicted by OCULAR-BEAM has a discontinuity in the middle of its prediction for the gold word `Surplus`. In contrast, the vertical offsets predicted by OCULAR-BEAM-SV at this location vary smoothly and more accurately match the true text baseline in the input image.

4.2 Font Detection Performance

We ran experiments with the italic font style model, OCULAR-BEAM-IT, on the Old Bailey test set (italics are infrequent in Trove). We evaluated the learned styles by measuring how accurately OCULAR-BEAM-IT was able to distinguish between italic and non-italic styles. Specifically, we computed the precision and recall for the system's predictions about which words were italicized. We found that, across the entire Old Bailey test set, OCULAR-BEAM-IT was able to detect which words were italicized with 93% precision at 74% recall, suggesting that the system did successfully learn both italic and non-italic styles.[4]

[4]While it seems plausible that learning italics could also improve transcription accuracy, we found that OCULAR-BEAM-IT actually performed slightly worse than OCULAR-BEAM. This negative result is possibly due to the extra difficulty of learning a larger number of font parameters.

System	CER	WER
Trove		
Google Tesseract	37.5	59.3
ABBYY FineReader	22.9	49.2
OCULAR (baseline)	14.9	33.0
OCULAR-BEAM	12.9	30.7
OCULAR-BEAM-SV	**11.2**	**25.6**
Old Bailey		
OCULAR (baseline)	14.9	30.8
OCULAR-BEAM	10.9	28.8
OCULAR-BEAM-SV	**10.3**	**27.5**

Table 1: We evaluate the output of each system on two test sets: Trove, a collection of historical newspapers, and Old Bailey, a collection of historical court proceedings. We report character error rate (CER) and word error rate (WER), macro-averaged across documents.

We can look at the typesetting layout predicted by OCULAR-BEAM-IT to gain insight into what has been learned by the model. The first line of Figure 3 shows the typesetting layout predicted by the OCULAR-BEAM-IT model for a line of a document image that contains italics. The second line of Figure 3 displays the corresponding portion of the input document image. From this example, it appears that the model has effectively learned separate glyph shapes for italic and non-italic versions of certain characters. For example, compare the template probabilities used to generate the `d`'s in `defraud` to the template probabilities used to generate the `d` in `hard`.

5 Conclusion

We began with an efficient simplification of the state-of-the-art historical OCR system of Berg-Kirkpatrick et al. (2013) and demonstrated two extensions to its underlying model. We saw an improvement in transcription quality as a result of removing a harmful independence assumption. This suggests that it may be worthwhile to consider still further extensions of the model, designed to more faithfully reflect the generative process that produced the input documents.

Acknowledgments

This work was supported by Grant IIS-1018733 from the National Science Foundation and also a National Science Foundation fellowship to the first author.

References

Kenning Arlitsch and John Herbert. 2004. Microfilm, paper, and OCR: Issues in newspaper digitization. the Utah digital newspapers program. *Microform & Imaging Review*.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2007. English Gigaword third edition. Linguistic Data Consortium, Catalog Number LDC2007T07.

Rose Holley. 2010. Trove: Innovation in access to information in Australia. *Ariadne*.

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*.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Machine translation: From real users to research*, pages 115–124. Springer.

Stephen Levinson. 1986. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech & Language*.

Slav Petrov, Aria Haghighi, and Dan Klein. 2008. Coarse-to-fine syntactic machine translation using language projections. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

Robert Shoemaker. 2005. Digital London: Creating a searchable web of interlinked sources on eighteenth century London. *Electronic Library and Information Systems*.

Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition*.

Hao Zhang and Daniel Gildea. 2008. Efficient multi-pass decoding for synchronous context free grammars. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.