Unsupervised Transcription of Historical Documents

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein
UC Berkeley
the prisoner at the bar. Jacob Lazarus and his wife, the prisoner, were both together when I received them. I sold eleven pair of them for three guineas, and delivered the remainder back to the prisoner. I sold seven pair of silk to Mark Simper: one pair of mixed, and two pair of thread to the footman, and one pair of thread to the barber.

Q. What is the footman’s name?

Frances Moses. I don’t know.

Henry Harris. I was standing at the Compter waiting for the sherriff’s officers to employ me: Moses’s daughter came for me to go and take the prisoner. I went to the Old Bailey
the prisoner at the bar. Jacob Lazarus and his wife, the prisoner, were both together when I received them. I fold eleven pair of them for three guineas, and delivered the remainder back to the prisoner. I fold seven pair of silk to Mark Simper: one pair of mixed, and two pair of thread to the footman, and one pair of thread to the barber.

Q: What is the footman’s name?
Frances Mofes. I don’t know.

Henry Harris. I was standing at the Compteh waiting for the sherriff’s officers to employ me: Mofes’s daughter came for me to go and take the prisoner. I went to the Old Bailey and Ch’: priftmer anhc bar. Jacob Lazarus and his IHP1 uh: prifoner. were both together when! rcccivcd lhczn. I fold eievén pair of than for xiirce quincas, and dclivrcd the rcll’l: in- d:r hack lo :1: prifuner. I fold ftvcm pair of filk to Mark Simpcr : nncpurir of mixcd. and. pair of Ifircad to lhz: foolnun, and or: pair of zhrzad to lh: barber. '

Q: What is the foolmarfs name?
Fraum Mgfr. I dun’: know.

Hairy Hzrvir. I was flandingar the Camp Icr waizin far the thcrrifs ufliceruco employ in: : Mo 3’: daughter came for me to 0 amf take the prifoncr. I Wm! to |hc Old aailcy
Unknown Fonts

positive
Unknown Fonts

long s glyph
Abstract
We present a generative probabilistic model, inspired by historical printing processes, for transcribing images of documents from the printing press era. By jointly modeling the text of the document and the noisy (but regular) process of rendering glyphs, our unsupervised system is able to decipher font structure and more accurately transcribe images into text. Overall, our system substantially outperforms state-of-the-art solutions for this task, achieving a 31% relative reduction in word error rate over the leading commercial system for historical transcription, and a 47% relative reduction over Tesseract, Google's open source OCR system.

1 Introduction
Standard techniques for transcribing modern documents do not work well on historical ones. For example, even state-of-the-art OCR systems produce word error rates of over 50% on the documents shown in Figure 1. Unsurprisingly, such error rates are too high for many research projects (Arlitsch and Herbert, 2004; Shoemaker, 2005; Holley, 2010). We present a new, generative model specialized to transcribing printing-press era documents. Our model is inspired by the underlying printing processes and is designed to capture the primary sources of variation and noise.

A key challenge is that the fonts used in historical documents are not standard (Shoemaker, 2005). For example, consider Figure 1a. The fonts are not irregular like handwriting – each occurrence of a given character type, e.g. 'a', will use the same underlying glyph. However, the exact glyphs are unknown. Some differences between fonts are minor, reflecting small variations in font design. Others are more severe, like the presence of the archaic long 's' character before 1804. To address the general problem of unknown fonts, our model learns the font in an unsupervised fashion. Font shape and character segmentation are tightly coupled, and so they are modeled jointly.

A second challenge with historical data is that the early typesetting process was noisy. Hand-carved blocks were somewhat uneven and often failed to sit evenly on the mechanical baseline. Figure 1b shows an example of the text's baseline moving up and down, with varying gaps between characters. To deal with these phenomena, our model incorporates random variables that specifically describe variations in vertical offset and horizontal spacing.

A third challenge is that the actual inking was also noisy. For example, in Figure 1c some characters are thick from over-inking while others are obscured by ink bleeds. To be robust to such rendering irregularities, our model captures both inking levels and pixel-level noise. Because the model is generative, we can also treat areas that are obscured by larger ink blotches as unobserved, and let the model predict the obscured text based on visual and linguistic context.

Our system, which we call Ocular, operates by fitting the model to each document in an unsupervised fashion. The system outperforms state-of-the-art baselines, giving a 47% relative error reduction over Google's open source Tesseract system, and giving a 31% relative error reduction over ABBYY's commercial FineReader system, which has been used in large-scale historical transcription projects (Holley, 2010).
Wandering Baseline
Wandering Baseline

Abstract

We present a generative probabilistic model, inspired by historical printing processes, for transcribing images of documents from the printing press era. By jointly modeling the text of the document and the noisy (but regular) process of rendering glyphs, our unsupervised system is able to decipher font structure and more accurately transcribe images into text. Overall, our system substantially outperforms state-of-the-art solutions for this task, achieving a 31% relative reduction in word error rate over the leading commercial system for historical transcription, and a 47% relative reduction over Tesseract, Google's open source OCR system.

1 Introduction

Standard techniques for transcribing modern documents do not work well on historical ones. For example, even state-of-the-art OCR systems produce word error rates of over 50% on the documents shown in Figure 1. Unsurprisingly, such error rates are too high for many research projects (Arlitsch and Herbert, 2004; Shoemaker, 2005; Holley, 2010). We present a new, generative model specialized to transcribing printing-press era documents. Our model is inspired by the underlying printing processes and is designed to capture the primary sources of variation and noise.

One key challenge is that the fonts used in historical documents are not standard (Shoemaker, 2005). For example, consider Figure 1a. The fonts are not irregular like handwriting – each occurrence of a given character type, e.g. 'a', will use the same underlying glyph. However, the exact glyphs are unknown. Some differences between fonts are minor, reflecting small variations in font design. Others are more severe, like the presence of the archaic long 's' character before 1804. To address the general problem of unknown fonts, our model learns the font in an unsupervised fashion. Font shape and character segmentation are tightly coupled, and so they are modeled jointly.

A second challenge with historical data is that the early typesetting process was noisy. Hand-carved blocks were somewhat uneven and often failed to sit evenly on the mechanical baseline. Figure 1b shows an example of the text's baseline moving up and down, with varying gaps between characters. To deal with these phenomena, our model incorporates random variables that specifically describe variations in vertical offset and horizontal spacing.

A third challenge is that the actual inking was also noisy. For example, in Figure 1c some characters are thick from over-inking while others are obscured by ink bleeds. To be robust to such rendering irregularities, our model captures both inking levels and pixel-level noise. Because the model is generative, we can also treat areas that are obscured by larger ink blotches as unobserved, and let the model predict the obscured text based on visual and linguistic context.

Our system, which we call Ocular, operates by fitting the model to each document in an unsupervised fashion. The system outperforms state-of-the-art baselines, giving a 47% relative error reduction over Google's open source Tesseract system, and giving a 31% relative error reduction over ABBYY's commercial FineReader system, which has been used in large-scale historical transcription projects (Holley, 2010).
Abstract

We present a generative probabilistic model, inspired by historical printing processes, for transcribing images of documents from the printing press era. By jointly modeling the text of the document and the noisy (but regular) process of rendering glyphs, our unsupervised system is able to decipher font structure and more accurately transcribe images into text. Overall, our system substantially outperforms state-of-the-art solutions for this task, achieving a 31% relative reduction in word error rate over the leading commercial system for historical transcription, and a 47% relative reduction over Tesseract, Google's open source OCR system.

1 Introduction

Standard techniques for transcribing modern documents do not work well on historical ones. For example, even state-of-the-art OCR systems produce word error rates of over 50% on the documents shown in Figure 1. Unsurprisingly, such error rates are too high for many research projects (Arlitsch and Herbert, 2004; Shoemaker, 2005; Holley, 2010). We present a new, generative model specialized to transcribing printing-press era documents. Our model is inspired by the underlying printing processes and is designed to capture the primary sources of variation and noise.

One key challenge is that the fonts used in historical documents are not standard (Shoemaker, 2005). For example, consider Figure 1a. The fonts are not irregular like handwriting – each occurrence of a given character type, e.g. a, will use the same underlying glyph. However, the exact glyphs are unknown. Some differences between fonts are minor, reflecting small variations in font design. Others are more severe, like the presence of the archaic long s character before 1804. To address the general problem of unknown fonts, our model learns the font in an unsupervised fashion. Font shape and character segmentation are tightly coupled, and so they are modeled jointly.

A second challenge with historical data is that the early typesetting process was noisy. Hand-carved blocks were somewhat uneven and often failed to sit evenly on the mechanical baseline. Figure 1b shows an example of the text's baseline moving up and down, with varying gaps between characters. To deal with these phenomena, our model incorporates random variables that specifically describe variations in vertical offset and horizontal spacing.

A third challenge is that the actual inking was also noisy. For example, in Figure 1c some characters are thick from over-inking while others are obscured by ink bleeds. To be robust to such rendering irregularities, our model captures both inking levels and pixel-level noise. Because the model is generative, we can also treat areas that are obscured by larger ink blotches as unobserved, and let the model predict the obscured text based on visual and linguistic context.

Our system, which we call Ocular, operates by fitting the model to each document in an unsupervised fashion. The system outperforms state-of-the-art baselines, giving a 47% relative error reduction over Google's open source Tesseract system, and giving a 31% relative error reduction over ABBYY's commercial FineReader system, which has been used in large-scale historical transcription projects (Holley, 2010).
The first dataset comes from a large set of images of the proceedings of the Old Bailey, a criminal court in London, spanning multiple lines. The values of such pixels that are not located in central text regions, or are treated as unobserved in the model since, more often than not, they are part of ink blots.

The Old Bailey dataset contains 10 documents, printed in consecutive decades. The first document is from 1715 and the last is from 1905. From the Old Bailey proceedings, we extracted a set of 20 documents into text. We will use these manual transcriptions to evaluate the output of our system. We choose the first document in each of the corresponding years, choose a random page in the document, and extracted an image of the first 30 consecutive lines of text consisting of full sentences.

The Old Bailey curatorial effort, after deciding to digitize Australian newspapers that were printed between 1700 and 1900 in England and Australia. Many of the images in historical collections are not already binary, and are part of large connected components of ink, often than not, they are part of ink blots. This is part of the reason that current OCR systems do not adequately handle such historical fonts, manually transcribed the images. For consistency, we binarized the images in our test sets that were not already binary.

We perform experiments on two historical datasets. Our second dataset is taken from a collection of digitized Australian newspapers that were printed in consecutive decades. The first document is from 1715 and the last is from 1905. Our second dataset is taken from a collection of newspapers, printed in consecutive decades. The first document is from 1715 and the last is from 1905. From the Old Bailey proceedings, we extracted a set of 20 documents into text. We will use these manual transcriptions to evaluate the output of our system. We choose the first document in each of the corresponding years, choose a random page in the document, and extracted an image of the first 30 consecutive lines of text consisting of full sentences.

Figure 6: Portions of several documents from our test set representing a range of difficulties are displayed. On document (a), which exhibits noisy typesetting, our system achieves a WER of 15.4. On document (c), which is relatively clean, we achieve a WER of 70.0. On document (d), which is severely degraded, we achieve a WER of 12.5. On document (b), which is also relatively clean, we achieve a WER of 12.5. On document (e), which is severely degraded, we achieve a WER of 70.0. On document (f), which is relatively clean, we achieve a WER of 12.5. On document (g), which is severely degraded, we achieve a WER of 70.0. On document (h), which is relatively clean, we achieve a WER of 12.5. On document (i), which is severely degraded, we achieve a WER of 70.0. On document (j), which is relatively clean, we achieve a WER of 12.5. On document (k), which is severely degraded, we achieve a WER of 70.0. On document (l), which is relatively clean, we achieve a WER of 12.5. On document (m), which is severely degraded, we achieve a WER of 70.0. On document (n), which is relatively clean, we achieve a WER of 12.5. On document (o), which is severely degraded, we achieve a WER of 70.0.
Our Approach
Generative Model

prisoner

r
Generative Model

Language Model

Typesetting Model

prison er
Generative Model

prisoner

Typesetting Model
Generative Model

prisoners

Typesetting Model
Generative Model

prisoner

Typesetting Model
Generative Model

p r i s o n e r

Rendering Model
Generative Model

Rendering Model

p r i s o n e r

prisoner
Generative Model

prisoner

Rendering Model
that the correspondence be regular) to learn the sequence of symbols, and we need to learn a cipherment (Ravi and Knight, 2008; Snyder et al., Kolak et al., 2003), but these do not deal directly specifically for correcting outputs of OCR systems. In the NLP community, generative models have been developed integrated typesetting models with language models, usually simultaneously. For example, the por-

A closely related area of work is automatic de-
mension. A wandering baseline

Kae and Nagy, 2000; Huang et al., 2006), which is not

of the challenges involved have been addressed

Relatively little prior work has built models specif-

Figure 2: An example image from a historical document (X): Typesetting Model

Wandering baseline

Historical font

Language Model

prison

3.1 Language Model

Our language model, is a Kneser-Ney 3-gram model (Kneser and Ney, 1995). We generate printed lines of text smoothed character sequences and let the line length

the character length of each line. We choose not to

that, formally, the model must separately generate generating an explicit stop character. This means

Ney, 1995). We generate printed lines of text

smoothed character

Our language model,

3.1 Language Model

vidual random variables with lower-case letters.

We let capital letters denote vectors of concate-

quer.

We take a generative modeling approach in-

spired by the overall structure of the historical
documents line by line; we present the generative

them jointly.

Generative Model

Type-setting Model

prison

3 Model

Taking

r

P(X|E,T)

X

prison

prish

X
Results

Old Bailey Court Proceedings (1715-1905)

Word Error Rate

- Google Tesseract
- ABBYY FineReader
- Ocular w/ NYT
- Ocular w/ OB
Jacob Lazarus and his wife, the prisoners were both together when I received them. I sold eleven pair of them for three guineas, and delivered the remainder back to the prisoner. I sold, seven pair of silk to Mark Simpert one pair of mixed, and two pair of thread to the footman, and one pair of thread to the barber.

Q: What is the footman's name?

Ms. What is the footman's name?

Franco Asyut, I don't know.

Nearly Norris. I was standing at the Computer waiting for the Sherrill's officers to employ me. A Moses's daughter came for me to go and take the prisoner. I went to the Old Bailey.
Conclusion

Thanks!