Letter Classification

Introduction

The goal of this study is to reconstruct a series of letters that John Mills wrote to his son, which he published under Letters of a Radio-Engineer to His Son in 1992 (New York: Harcourt, Brace and Company). The letters are in need of reconstruction because Mr. Mill's typists, Typist-1 and Typist-2, transcribed the letters incorrectly.

We know that Typist-1 typed letters 1, 8 and 16 and Typist-2 typed letters 4,9 and 18. We attempt to attribute the letters to their respective typists by "training" models based on our prior information on who typed a particular letter. Training amounts to informing the model on what kinds of errors each of the typist s most frequently commit. With this knowledge, when we give our model the original letter and a transcribed letter, the model can make an informed guess on who could have typed the transcribed letter, based on the most frequent types of errors present in the transcribed letter.

We will begin by training on a single letter for each typist, and reporting the resulting evidence (in decibels) of our attribution compared to the known author of the letter. We will do this for every possible combination of letters available (36 in total). We will then perform this same analysis by first training on two letters by each typist, for every combination of letters available (18 in total). These results will hopefully show the reliability of our method. Finally, we will train on all three available letters, and attempt to attribute the letters with unknown transcribers.

Calculating Evidence

Jaynes defines *evidence* (which we shall refer to as Jayne's evidence henceforth) as $e(H|DX) \equiv 10\ log_{10}\ O(H|DX)$[1]. Notice that this definition of *evidence* is not the same as the evidence from the Bayes Rule. Jayne's evidence can be expressed as $e(H|DX) \equiv e(H|X) + 10\ log_{10} \frac{P(D|HX)}{P(D|\bar{H}X)}$[2].

In this context, prior is the belief that a unclassified letter was written by a particular typist before we evaluate the unclassified letter. For all the predictions that we conduct in this study, we train each typist by the same number of letters. Thus, by the **Rule of Succession**, we assign the same probability for the letter being typed by Typist-1 and the letter being typed by Typist-2. This implies that

$$e(H|X) = 10log_{10}O(H|X) = 10log_{10}\frac{P(H|X)}{P(\bar{H}|X)} = 0$$

.

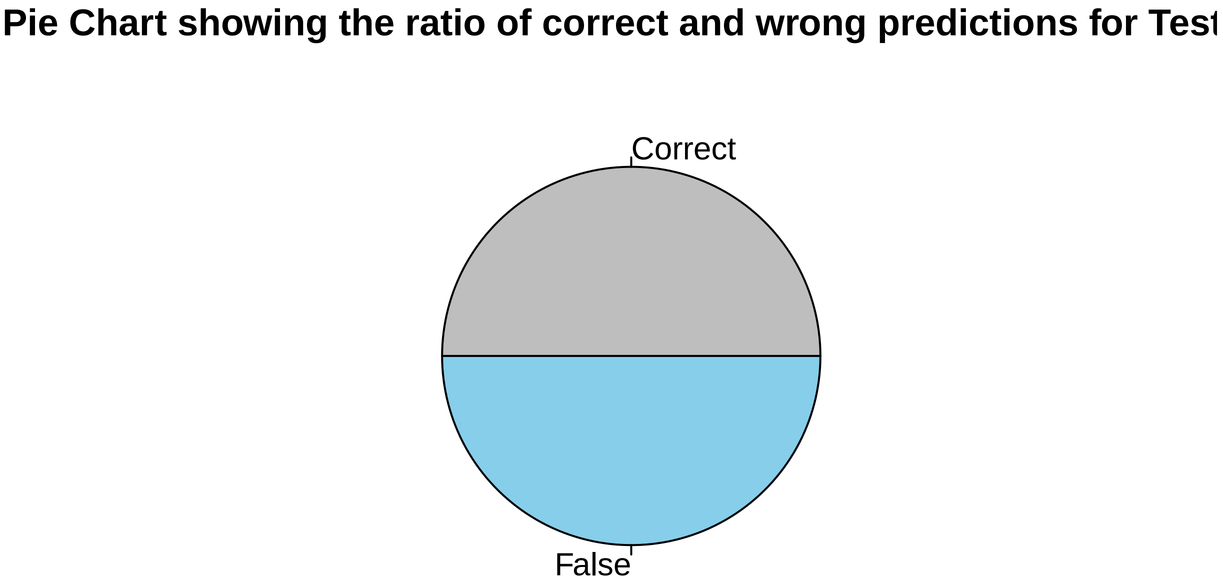
Test A : Training on a single letter

There are 3 ways to select a letter typed by Typist-1 and 3 ways to select a letter typed by Typist-2. Selecting 2 letters from 6 letters leaves us with 4 letters. For each of the 4 remaining letters, we use the trained model to predict who typed it. This means that we can run (3 * 3 * 4 = 36) tests. For each of the predictions, we exactly know who typed the letter. So, we can objectively know whether the prediction made by the model was right or wrong. Depending on the percentage of correct predictions, we can decide on how confident we can be in our methodology.

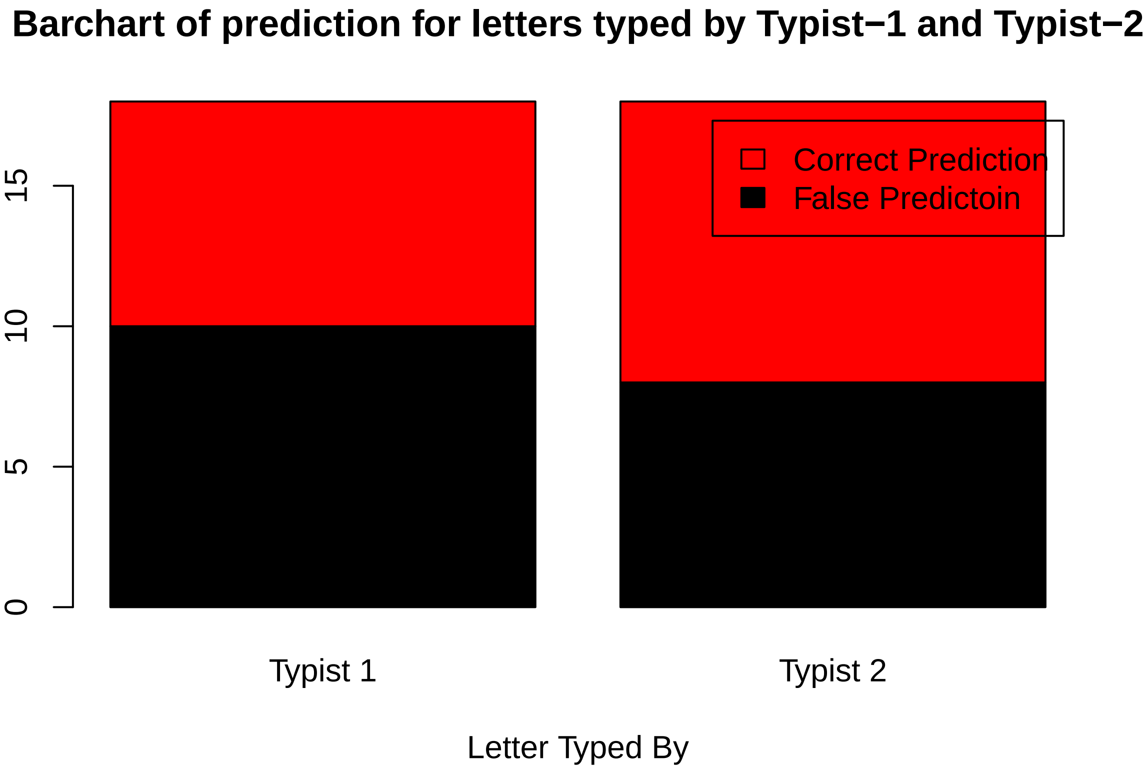
[1] Jaynes, E. T. Probability theory: the logic of science. Cambridge University Press, 2003, p.91.(This is referred to as equation 4.8 in the book)

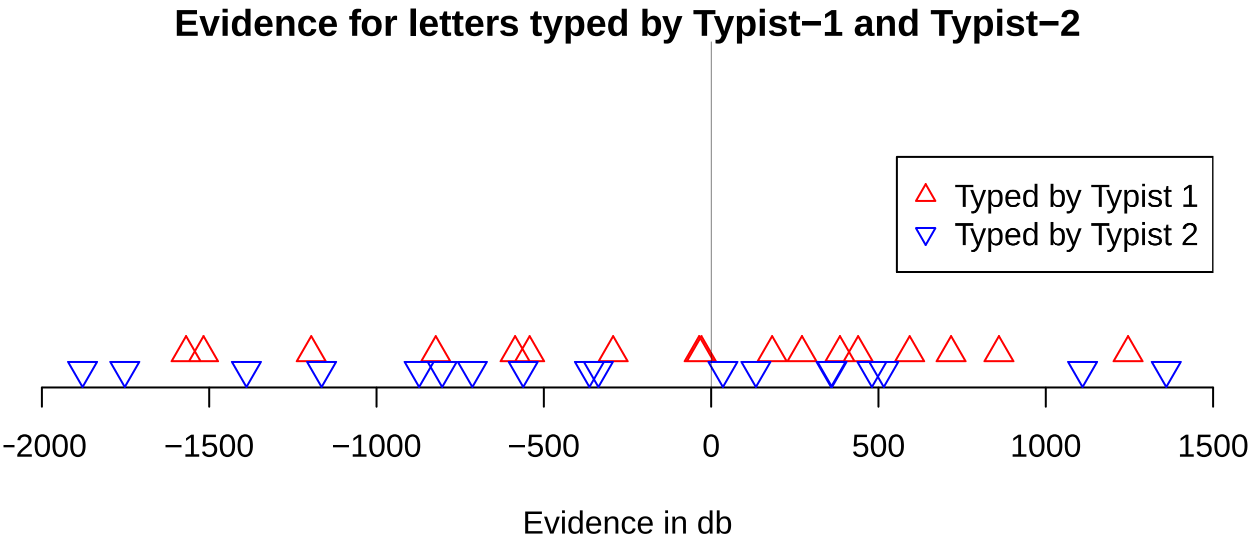
[2] Jaynes, E. T. Probability theory: the logic of science. Cambridge University Press, 2003, p.91. (This is referred as equation 4.9 in the book)

Results



Out of the 36 predictions, only 18 of them were correct. Of the 18 letters typed by Typist-1, 8 of them were correctly predicted whereas of the 18 letters typed by Typist-2, 10 of them were correctly predicted. The predictions from training using a single letter for Typist-1 and Typist-2 performed as well as randomly flipping a coin; the expected value of predicting correctly was only 50%.



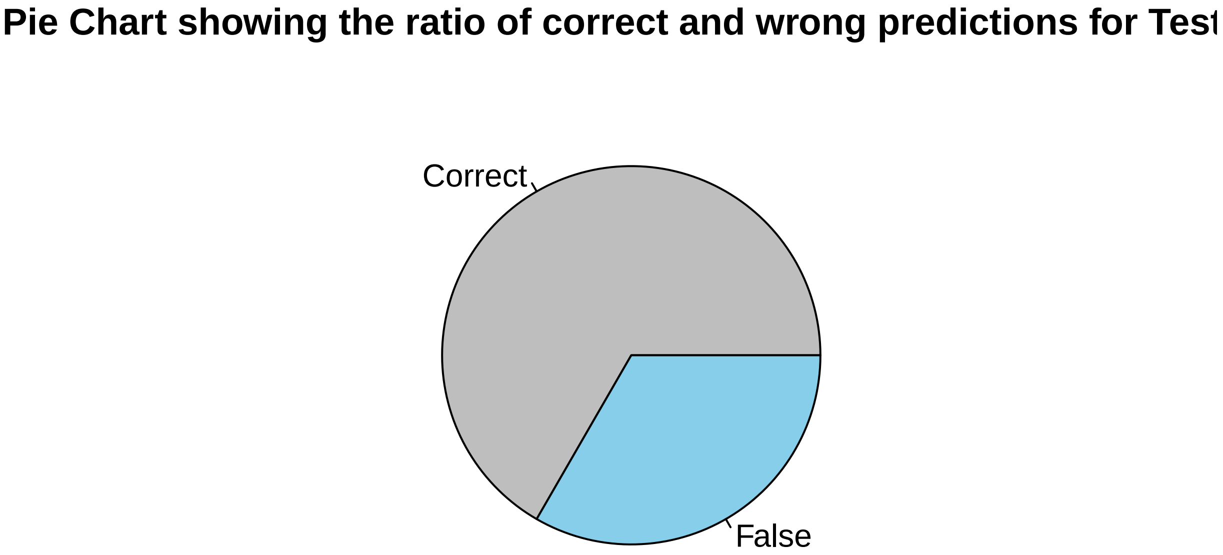

Evidence for letters typed by Typist−1 and Typist−2

Test B: Training on two letters

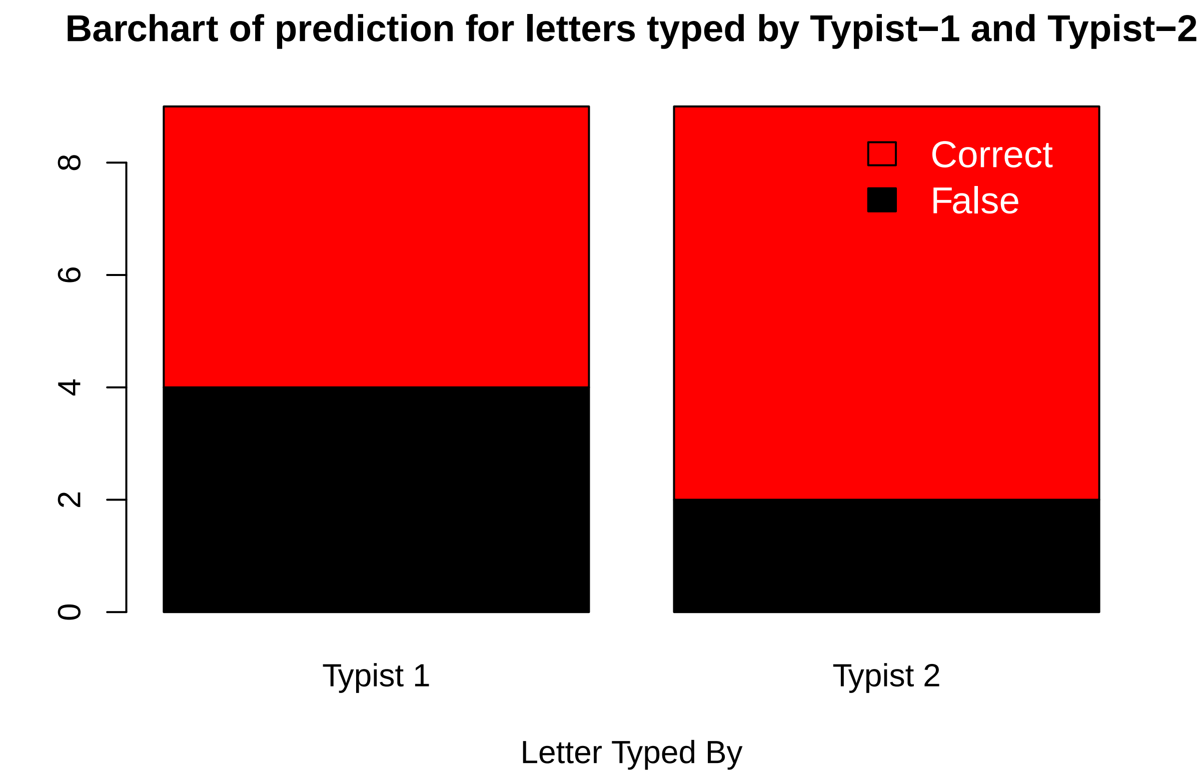
Using the same letters as in our prior analysis, we ran new tests in which we trained using two letters from each typist attempting to attribute the one remaining, allowing us to perform (3 * 3 * 2 = 18) tests. Just as before, we know who typed each letter and we have the correct transcript of the original copy, so we are able to report on our accuracy.

Results

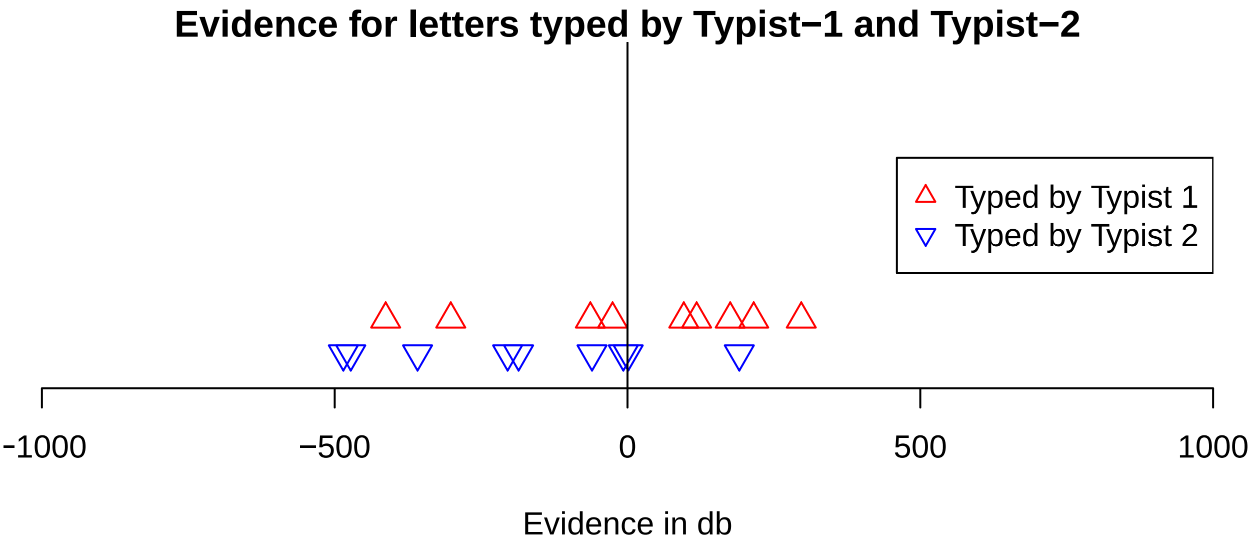
This set of predictions were more accurate compared to Test A. Of the 18 predictions made in Test B, 12 of them were correct, yielding an accuracy of ~ 67%. This gives us confidence in our methodology as unlike Test A, it performs better than random chance.


Pie Chart showing the ratio of correct and wrong predictions for Tes

The wrong predictions of letters did not come from evenly from Typist-1 and Typist-2. Of the 6 letters that were incorrectly classified, 4 of them were typed by Typist-1 whereas 2 only 2 of them were typed by Typist-2.



After training each typist by 2 letters, the expected value of making an incorrect prediction is $\frac{1}{3}$. This is still a very large number. However, the jump in accuracy from one letter to two makes us confident that by training on three letters, we will have a much stronger accuracy.



Training on three letters

Unlike in our previous tests, for these tests we are training on all three of our known letters by each typist, and attempting to attribute the remaining unattributed letters for which we have the originals. The unattributed letters that we have originals for include letters 3, 7, 10, 11, 15 and 22. For all the other unattributed letters we do not possess original copies, and therefore are unable to provide evidence for them.

Letter	Evidence	Attribution
3	-185.14701	Typist 2
7	211.00767	Typist 1
10	-28.69143	Typist 2
11	73.30541	Typist 1
15	-105.61027	Typist 2
22	253.50924	Typist 1

Conclusions

As shown in the results of our final tests, we have attributed letters 7, 11 and 22 to Typist-1, and letters 3, 15 and 10 to Typist-2. Since the accuracy of Test B was about 67%, we are at confident that this classification will perform better than random chance. It is reasonable to assume that training with 3 letters will perform even better than training by by only 2 letters. We expect the accuracy of our test to be around 75%.

Analyzing the magnitude of evidence for each letter, we are least confident in our attribution of letter 10 as it has the lowest magnitude of evidence at around 28. This still means that as per our analysis, we should be 280 times more likely to believe that letter 10 was typed by Typist-2 as compared to Typist-1. If any, we suspect that we incorrectly classified letter 10 as the evidence for all the other classifications are pretty high.

Without more letters to train on, this is the best method we can conceive of to attribute these letters. We are confident that our methodology of classifications will get better once we get possession of more letters to train on.

Appendix

Table for training on a single letter

Train.Typist.A	Train.Typist.B	Test.Letter	Letter.Typed.By	Evidence	Correct
1	4	18	2	-872.66954	TRUE
1	4	9	2	-803.76117	TRUE
1	4	16	1	-822.96330	FALSE
1	4	8	1	-542.40943	FALSE
1	9	18	2	-1752.30766	TRUE
1	9	4	2	-1164.16319	TRUE
1	9	16	1	-1569.10026	FALSE
1	9	8	1	-1195.08542	FALSE
1	18	9	2	-1878.51144	TRUE
1	18	4	2	-1388.78774	TRUE
1	18	16	1	-2125.80817	FALSE
1	18	8	1	-1516.92798	FALSE
8	4	18	2	515.87280	FALSE
8	4	9	2	361.18600	FALSE
8	4	16	1	717.00819	TRUE
8	4	1	1	384.91127	TRUE
8	9	18	2	-363.76532	TRUE
8	9	4	2	-336.96102	TRUE
8	9	16	1	-29.12877	FALSE
8	9	1	1	-35.84846	FALSE
8	18	9	2	-713.56427	TRUE
8	18	4	2	-561.58556	TRUE
8	18	16	1	-585.83668	FALSE
8	18	1	1	-292.84132	FALSE
16	4	18	2	1360.08838	FALSE
16	4	9	2	1109.94312	FALSE
16	4	8	1	1245.93590	TRUE
16	4	1	1	860.01573	TRUE
16	9	18	2	480.45026	FALSE
16	9	4	2	358.34003	FALSE
16	9	8	1	593.25991	TRUE
16	9	1	1	439.25601	TRUE
16	18	9	2	35.19285	FALSE
16	18	4	2	133.71549	FALSE
16	18	8	1	271.41735	TRUE
16	18	1	1	182.26314	TRUE

Table for training on two letters

Train.A.1	Train.A.2	Train.B.1	Train.B.2	Test.Letter	Letter.Typed.By	Evidence	Correct
1	16	4	9	8	1	215.497580	TRUE
1	16	4	18	8	1	96.136097	TRUE
1	16	9	18	8	1	-25.522489	FALSE
8	1	4	9	16	1	-63.378424	FALSE
8	1	4	18	16	1	-301.783619	FALSE
8	1	9	18	16	1	-412.977415	FALSE
8	16	4	9	1	1	296.768784	TRUE

Train.A.1	Train.A.2	Train.B.1	Train.B.2	Test.Letter	Letter.Typed.By	Evidence	Correct
8	16	4	18	1	1	175.407473	TRUE
8	16	9	18	1	1	118.011109	TRUE
1	16	4	9	18	2	1.207651	FALSE
1	16	4	18	9	2	-185.993292	TRUE
1	16	9	18	4	2	-204.712519	TRUE
8	1	4	9	18	2	-358.505900	TRUE
8	1	4	18	9	2	-472.557773	TRUE
8	1	9	18	4	2	-485.242239	TRUE
8	16	4	9	18	2	190.933284	FALSE
8	16	4	18	9	2	-7.115255	TRUE
8	16	9	18	4	2	-60.582567	TRUE