A Mathematical Explanation of Burrows's Delta *

Sterling Stein **Shlomo Argamon**
Linguistic Cognition Laboratory
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616

1 Introduction

While many methods have been applied to the problem of automated authorship attribution, John F. Burrows's "Delta Method" [1] is a particularly simple, yet effective, one [2, 3]. The goal is to automatically determine, based on a set of known training documents labeled by their authors, who the most likely author is for an unlabeled test document. The Delta method uses the most frequent words in the training corpus as the features that it uses to make these judgments. The Delta measure is defined as:

> the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text [4].

The Delta of the test document is computed with respect to each of the training documents, and that author whose training document has minimal Delta with the test document is chosen for attribution.

While this method is intuitively reasonable, we may still ask: Why does this method work so well? Why z-scores? Why mean of absolute differences? Perhaps if we understood the mathematical underpinnings of the method, we could modify it to make it more effective. Furthermore, we would know better when it is applicable and when using it would not make sense.

Hoover has implemented Delta-based attribution as a spreadsheet [5]; we have re-implemented it as Java program with several options for using different variations, which we are experimenting with. This program will be made available to the research community once it has reached a stable state.

*This document appears in Digital Humanities ALLC 2006

2 Probabilistic formulation

This section will briefly show how Burrows's Delta may be profitably viewed as a method for ranking authorship candidates by their probability. Let X and Y be n-dimensional vectors of the word frequencies in two documents. Note that the z-score is obtained by subtracting out the mean and dividing out the standard deviation. Then the Delta measure between these documents can be reformulated:

$$\begin{aligned}\sum_{i=1}^{n} |z(X_i) - z(Y_i)| &= \sum_{i=1}^{n} \left| \frac{X_i-\mu_i}{\sigma_i} - \frac{Y_i-\mu_i}{\sigma_i} \right| \\ &= \sum_{i=1}^{n} \left| \frac{(X_i-\mu_i)-(Y_i-\mu_i)}{\sigma_i} \right| \\ &= \sum_{i=1}^{n} \left| \frac{X_i-Y_i}{\sigma_i} \right|\end{aligned}$$

Note that the value of μ_i, the mean frequency of word i, cancels out, with the only effect of the training corpus as a whole being the normalizing factor of σ_i, the standard deviation for word i.

Thus, Delta is like a scaled distance between the 2 documents. It is not the ordinary distance "as the crow flies", but rather it is the sum of each dimension independently, called the *Manhattan distance*. It is like walking the streets of Manhattan as we stay on the grid.

Note that if we consider the mean of a distribution in place of Y_i, this has a form similar to a Laplace probability distribution [6]. Specifically, it is the exponent of the product of independent Laplace distributions. Thus, we are assuming that the individual document that we are comparing the testing document against is a sort of average document for that author. Taking of the z-score corresponds to the normalization in the exponent. So, in a sense, Delta is measuring the probability of a document being written by an author taking each word

frequency independently and then choosing the document with the highest probability.

In effect, that we are using the z-score means that we are estimating the parameters of the Laplace distribution by the sample mean and standard deviation. However, the maximum likelihood estimator of the Laplace distribution is the median and the mean absolute deviation from the median [2, 3]. This gives us our first variation of the Delta measure. Instead of using the z-score, we should use the median and "median deviation". Whereas the Delta measure gives a distance in a purely abstract space, this variation provides a well-founded probability.

Now that we know we are looking at a probabilistic model, we can try putting in other distributions. A commonly-used distribution is the Gaussian, or normal, distribution. It is similar to the Laplace distribution except that it uses a sum of squares rather than of absolute values, based on the mean of the mean and standard deviation, hence using the z-score *is* appropriate here. Note that in this case, we have the "as the crow flies" Euclidean distance instead of the Manhattan distance.

Further, note that the previous measures consider each dimension independently. In this sense, they are axis-aligned. This means that the use of each word is assumed to have nothing to do with the use of any other words. Of course, this assumption is false, but may be a reasonable approximation. To take this co-occurrence into account, we can use a rotated method, eigenvalue decomposition. Previously we used the z-score of individual words. Instead of using the standard deviation, we can generalize to using the entire covariance matrix. In this, we take the largest magnitude eigenvalues from the covariance matrix and use the corresponding eigenvectors as the features.

3 Evaluation

We are currently performing empirical tests. To compare these new variants to the original and each other, we will use 3 corpora. First, we will use the data in the spreadsheets from [5] to check that our implementation is working properly and so we can directly compare results. The second will be a collection of essays written by students taking a psychology course. There are up to 4 essays by each author. The third will be the 20 newsgroups corpus from `http://people.csail.mit.edu/jrennie/20Newsgroups/`. These corpora will be split into testing and training such that the training has only 1 or 0 of each author. Having 0 allows for the possibility of an unknown author. Each of these variations will be run on the corpora and the results of each classification will be split into 5 categories, based on who the attributed author is (lowest Delta candidate), and whether the attribution is considered reliable. This is determined via a threshold on attribution confidence, such that confidences below the threshold are considered "unknown". To determine the confidence, we use Delta-z, following Hoover [3].

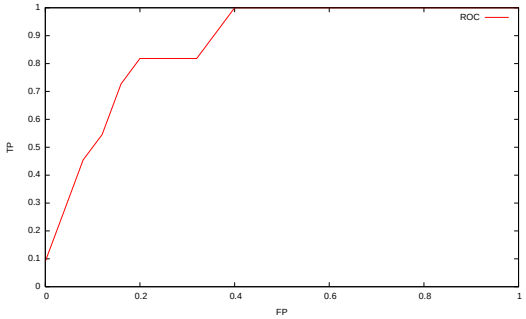

Figure 1: A sample ROC curve. The X-axis is the false positive rate FP, while the Y-axis is the true positive rate TP. Note that TP increases monotonically with FP, and that TP=1 once FP=0.4.

Say/Is	A	B	U
A	α	β	γ
B	β	α	γ
U	δ	δ	ϵ

The first type of classification decision, α, is a true positive, where the correct author is attributed. Next, β is a false positive, where a known author is chosen, but not the correct one. Another false positive is γ, where a known author is chosen, but the true author is not in the training. Fourth, δ is a false negative, where no author was recognized, but should have been. Finally, ϵ is a true negative, where the true author was not in the training set and was not recognized.

These allow us to calculate the true positive and false positive rate, where:

$$true\ positive\ rate = \alpha/(\alpha + \delta)$$

$$false\ positive\ rate = (\beta + \gamma)/(\beta + \gamma + \epsilon)$$

These values can be used to make a receiver operating characteristic (ROC) graph (Figure 1), which

shows the sensitivity of a method to the signal in the data. It is trivial to declare everything a negative, which would give 0 to both TP and FP. As you classify more instances as positive, there is more of a risk that an instance classified as positive is not. An overall measure of a method's efficacy can be computed as the area under the ROC curve (AUC); this score will be used for comparison between the methods. AUC measures the trade off between false positives and false negatives, with the baseline at 50% (where the line goes straight from (0,0) to (1,1)) and the best possible value of 1, meaning always getting true positives with no false positives. In this way, it will allow us to judge and compare how well the different variations work.

4 Conclusion

We have reformulated Burrows's Delta method in terms of probability distributions. This allows us to extend the method to use multiple different probability distributions and to interpret the result as a probability. At the conference, we will present results comparing the effectiveness of these variations. More importantly, this work provides a more solid foundation for understanding Delta. In particular, the probabilistic assumptions that it makes, such as word frequency independence and that authors have similarly-shaped word-frequency distributions, are made explicit, allowing us better understanding of the uses and limitations of the method. For example, it is now clear why the method should only be applied to documents all of the same well-defined text type.

References

[1] J. F. Burrows, "Delta: a measure of stylistic difference and a guide to likely authorship," *Literary and Linguistic Computing 17*, pp. 267–287, 2002a.

[2] D. Hoover, "Delta prime?," *Literary and Linguistic Computing 19.4*, pp. 477–495, 2004b.

[3] D. Hoover, "Testing burrows's delta," *Literary and Linguistic Computing 19.4*, pp. 453–475, 2004a.

[4] J. Burrows, "The englishing of juvenal: Computational stylistics and translated texts," *Style 36*, pp. 677–699, 2002.

[5] D. Hoover, "The delta spreadsheet," *Literary and Linguistic Computing 20*, 2005.

[6] S. K.-M. J. Higgins, *Concepts in Probability and Stochastic Modeling*. Duxbury Press, 1 ed., 1994.

[7] P. Juola, "A prototype for authorship attribution software," *Literary and Linguistic Computing 20*, 2005.