Optimum parameter selection for K.L.D. based Authorship Attribution for Gujarati

Parth Mehta
DA-IICT, Gandhinagar
parth.mehta126@gmail.com

Prasenjit Majumder
DA-IICT, Gandhinagar
prasenjit.majumder@gmail.com

Abstract

We examine several quantitative techniques of authorship attribution that have gained importance over the time and compare them with the current state of the art Z-score based technique. In this paper we show how comparable the existing techniques can be to the Z-score based method, simply by tuning the parameters. We try to find the optimum values for number of terms, smoothing parameter value and the minimum number of texts required for creating an author profile.

1 Introduction

Authorship attribution and author profiling have a long standing history dating back to 19th century (Mosteller and Wallace, 1964). While authorship attribution deals with determining whether or not a given author has written the given article, author profiling aims at determining the age, gender, education level, etc. of the author form his article (Koppel et al., 2002). In the current work we focus only on authorship attribution. Authorship attribution techniques can be broadly classified into *linguistic techniques* and *Statistical techniques*. Until early 90's majority of the work done was from a linguistic perspective. Only after (Holmes, 1994) did statistical methods gain importance. There were many attempts to solve this particular problem using various statistics of texts like mean sentence length, term frequency distributions (Zipf, 1932), word lengths, character frequencies (Peng et al., 2003), vocabulary richness, etc. Rudman (1998) proposed nearly 1000 different measures that could be used. A whole new field of study called stylometry evolved from these type of studies. In 2002, (Burrows, 2002) proposed a novel technique based on Z-score that covered many of the above features, specially vocabulary difference and difference in term distribution into a single measure. Later, Savoy(2010) modified this Z-Score that improved the result drastically and this technique is currently the state of the art for authorship attribution. In this paper we compare three major statistical techniques for authorship attribution to the state of the art Z-score based technique.

2 Corpus Details

The absence of a replicable and reliable corpora daunts the field of authorship attribution and more so for Indian languages. To the best of our knowledge the only work available to date for Indian languages is by (Bagavandas and Manimannan, 2008) where the corpus consisted of 55 Tamil articles, 32 of which were attributed and 23 disputed and there were only three possible authors. Having this fact in mind and following the steps of (Savoy, 2012) the authors developed a corpora consisting of 5039 newspaper articles from the newspaper *Gujarat Samachar*. These articles consist of 49 different weekly articles, from the supplements *Shatdal* and *Ravi Purti*, written by 40 distinct authors in the time period of 01-Dec-2011 to 1-Dec-2012 and is made available on our website[1], alongwith the details of articles and authors. These articles span various categories like Science and Technology, short stories, Health and Fitness, Politics, etc. Though our corpus is more biased towards fiction(short stories & novels) this should not affect the performance because, unlike categories like health or art, stories seldom have a large overlap of vocabulary. Mean length of the documents was found to be 972 (Minimum: 85 Maximum: 1774, Median: 909,Standard deviation: 382). These texts were all available in the standard UTF-8 format and hence the only pre-processing that we did was to remove the punctuations, numerals and author names from the text. Except this no other pre-

[1] http://irlab.daiict.ac.in/tools.php

International Joint Conference on Natural Language Processing, pages 1102–1106,
Nagoya, Japan, 14-18 October 2013.

processing was done, each morphological variant was treated as a unique term and also there was no word sense disambiguation to distinguish between same words having different meaning. Concept of capitalization does not exist as such in Gujarati language. Our experiments being completely statistical in nature can be replicated very easily without any knowledge of Gujarati language.

3 Experiment details

We compare four different Authorship attribution methods mentioned in (Savoy, 2012), namely Delta method, Chi-squared method, Z-Score based method and Kullback Leibler divergence based method. Our aim is to examine whether or not tuning the parameters of K.L.D. based method can produce results comparable to the state of the art Z-score based method. In this section we briefly describe each of the four methods and then explain the parameters that can affect K.L.D. based authorship attribution. All these methods are profile based methods i.e. for each of the N authors we create an author profile A_k where $k \in \{1, ..., N\}$. These profiles are created from the documents for which the true author is already known. Disputed document Q is then compared to each author profile A_k using a metric $D(Q, A_k)$ and is attributed to that author for whom D is minimum. In other words for given query text Q and author profiles A_k

$$A_{correct} = \underset{k \in N}{\operatorname{argmin}}\{D(Q, A_k)\} \qquad (1)$$

The distance function D depends on the method used and is defined separately for each method and so is true for the author profile A_k.

The parameters in these experiments that are to be set heuristically include the value of the smoothing technique and smoothing parameter (λ) for that technique, the minimum number of texts(N) that have to be used in order to create a reasonably good author profile and the number of terms(X) considered to create the author and document profiles. Due to several studies readily available, we directly use Lidstone smoothing technique without further experimentation. Our main aim is to find the optimum value of these parameters for a corpus of Gujarati articles.

3.1 Delta Method

Delta method was first proposed by (Savoy, 2012). It uses a term-document index along with Z-score defined by equation 2 below

$$Z_{score}(t_{ij}) = \frac{tf_{ij} - mean_i}{sd_i} \qquad (2)$$

Z-score is calculated for each term t_{ij} where $i \in \{1, ..., T\}$ and $j \in \{1, ..., M\}$. T and M are the total number of unique terms and total number of documents in the corpus respectively. tf_{ij} is the term frequency of term i in document j, $mean_i$ and sd_i are the mean and standard deviation of frequency of term t_i in the entire corpus. Using this we can represent each document as a vector of Z-scores for each of its terms. Hence each document can be represented as a vector $d_j = [Z_{score}(t_{1j}), Z_{score}(t_{2j}), ..., Z_{score}(t_{mj})]$ for a particular value of j. Having this representation for each document an author profile A_k can then be created by taking the mean of these vectors for all the articles known to be written by that particular author.

Next we represent the query text Q in the same manner, as a vector of Z-scores. We then find the author profile that is closest to Q using equation 1, the distance function being defined as below.

$$D_1(Q, A_j) = \frac{1}{T}.\sum_{i=1}^{T} |Z_{score}(t_{iq}) - Z_{score}(t_{ij})|$$

t_{iq} denotes term t_i in query text, and t_{ij} denotes term t_i in author profile j.

3.2 Chi-Squared distance based method

Chi-Squared distance based method is based on the well known Pearson's χ^2 test, used to compute the similarity between two distributions. In this method a document is represented as a vector $d_j = [p(t_{1j}), p(t_{2j}), ..., p(t_{Mj})]$, where $p(t_{ij})$ is normalised frequency of term t_i in a given document j. Author profile A_k is prepared by first combining all the documents pertaining to a particular author k, and then calculating the normalised frequency for this combined document. Considering Q and A_k as term distributions we can now use χ^2distance to find the degree of similarity between the two. The distance function in this case is as shown below

$$D_2(Q, A_k) = \sum_{i=1}^{T} \frac{(q(t_i) - a_k(t_i))^2}{a_k(t_i)}$$

where $q(t_i)$ is the normalised frequency for term t_i the query text Q and $a_k(t_i)$ is that for the k^{th} author profile.

3.3 Z-Score based method

This method is currently the state of the art method for authorship attribution using quantitative analysis. It was proposed by Savoy (2012) and is a modification of the Delta method mentioned in section 3.1. One of the two major modifications is the method of calculating Z-Score. Savoy (2012) proposed using the expected value of term frequency and the expected standard deviation compared to the sample mean and sample standard deviation that were used in Delta method. So in this case any term t_{ij}, i^{th} term in j^{th} document, is considered to be drawn from a binomial distribution with parameters $n0$ and $p(t_i)$. $n0$ is the length of the document for which t_{ij} is to be estimated and $p(t_i)$ is the probability of term t_i occurring in the entire corpus. Hence the expected value for t_{ij} is $n0.p(t_i)$ and the expected standard deviation is $\sqrt{n0.p(t_i).(1-p(t_i))}$. The modified Z-score can then be calculated as

$$Z^*_{score}(t_{ij}) = \frac{tf_{ij} - n_0.p(t_i)}{\sqrt{n_0.p(t_i).(1-p(t_i))}} \quad (3)$$

This Z-Score can than be used in the same way as used in Delta method. Another change in this method as compared to the Delta method is the distance function used.

$$D_3(Q, A_j) = \frac{1}{T}.\sum_{i=1}^{T}\left(Z^*_{score}(t_{iq}) - Z^*_{score}(t_{ij})\right)^2$$

where t_{iq} denotes term t_i in query text, and t_{ij} denotes term t_i in author profile j.

3.4 K.L.D. based method

K.L.D. based method is somewhat similar to the Chi-squared distance method in that this method also looks upon normalised word frequencies as a probability distribution. The author profiles and document profiles in this case are exactly the same as that in the Chi-squared distance based method. Kullback Leibler Divergence between the two probability distributions, namely author profile A_k and query text Q is defined as below

$$D_{KL}(Q||A_k) = \sum_{i=1}^{T} \ln\left(\frac{a_k(t_i)}{q(t_i)}\right) q(t_i)$$

where $q(t_i)$ is the normalised frequency for term t_i the query text Q and $a_k(t_i)$ is that for the k^{th} author profile. Author with profile A_k with minimum divergence from Q is considered to be the author for the disputed text.

4 Results and Evaluation

In this section we present the results of applying these four aforementioned techniques on our corpus. We also include one more technique apart from these four in which we use Delta method albeit with distance function D_3. We use the same evaluation strategy used by Savoy (2012). At a time we choose one article to be the disputed text Q and use all other articles to create the author profiles A_k. This is repeated for every article present in the corpus. Accuracy is then calculated in two ways: by finding the total number of articles attributed correctly irrespective of the authors (micro average) and by finding the accuracy for each author individually and then defining the overall accuracy as the average of these individual values (macro average). While experimenting with the number of texts required to create an author profile, for each article we select p articles from each author to create the author profiles. The concept of macro and micro average remain the same. But since we are selecting these p articles randomly, we perform a 10-fold cross validation to assure statistically significant results. In this case we report mean accuracy. Table 1 below shows the result for using different values of λ, with X and N remaining constant. All the terms with $tf > 10$ and $df > 2$, were considered for the Z_{score} and $K.L.D.$ based approaches while for Delta method top 400 terms were considered. For the chi-square based method the condition $tf > 2$ was used. All these conditions are based on the best performing parameter value as found by (Savoy, 2012) and hence would make a good starting point. Above this we consider only those terms that belong to at least two author profiles so as not to make the task trivial. The size of the training set for this experiment was $N = N_{max}$ i.e. all the available articles (except the query text Q) are used to create the author profile. For each experiment the best performing parameter value is considered to be the baseline and other values are compared against them for statistically significant difference, using a two sided sign test.

Method	Parameter	Micro-Average	Macro-Average
Z-Score	$\lambda = 0$	86.14%	87.38%†
	$\lambda = 0.1$	**88.73%**	**90.45%**
Delta (D_1)	$\lambda = 0$	26.10%†	24.69%†
Delta (D_3)	$\lambda = 0$	84.24%†	86.00%†
KLD	$\lambda = 0.01$	77.17%†	70.38%†
	$\lambda = 0.001$	88.57%	85.44%†
χ^2 Method	$\lambda = 0$	12.15%†	14.73%†

Table 1: Effect of variation in λ

† Significant performance difference ($\alpha = 1\%$, two-sided sign test)

For further experiments we consider only the best performing value of the smoothing parameter and show that with proper feature selection, $i.e.$ by selecting proper number of terms, K.L.D. based approach can give results comparable to the state of the art Z-score based approach. Chi-squared method and Delta method(using D_1 distance) perform poorly and hence we do not consider them in further experimentation. All further experiments are performed only on Z-Score based method, Delta method(using D_3 distance) and K.L.D. based method.

Method	Parameter	Micro-Avg	Macro-Avg
Z-Score	$tf > 10, df > 3$	88.73%	90.45%†
	$tf > 100, df > 3$	84.33%†	86.45%†
Delta (D_3)	Top 100 terms	76.10%†	74.69%†
	Top 400 terms	84.24%†	86.00%†
KLD	$tf > 10, df > 3$	88.57%†	85.44%†
	$tf > 100, df > 3$	90.55%	88.75%
	$tf > 1000, df > 3$	**91.35%**	**91.73** %

Table 2: Effect of variation in X

† Significant performance difference ($\alpha = 1\%$, two-sided sign test)

Table 2 shows the variation in performance of these methods when the number of terms are varied. For Z-score based method and K.L.D. based method we choose terms based on their term frequencies in the corpus. We keep document frequency constant because increasing it would lead to selection of only those terms which are prevalent across more number of documents. These terms will make the author profiles less distinguishable and result in poor overall performance. For Delta method fewer number of terms always perform better (Burrows, 2002). Hence we use 100 and 400 terms respectively as done by (Burrows, 2002) and followed by (Savoy, 2012)

Further we investigate the effect of reducing the training set $i.e.$ the number of texts used to create author profile. For this we select the smoothing parameter and the number of terms that performed best in the previous two experiments. For Z-Score based method we use the criteria $tf > 10, df > 3$, for K.L.D. based method we use $tf > 1000, df > 3$ and for delta method we use top 400 most frequent terms. Table 3 shows the performance of the three systems as we vary the size of training set. N_{max} refers to the maximum number of articles that can be used to create the author profiles. In our case it is $N_k - 1$, where N_k is the total number of documents for the K^{th} author. Clearly when the size of the training set is small K.L.D. based method fares much better than all other techniques.

Method	Parameter	Micro-Average	Macro-Average
Z-Score	$N = 10$	52.14%†	54.17%†
	$N = 40$	82.39%†	84.45%†
	$N = N_{max}$	88.73%	90.45%
Delta	$N = 10$	22.10%†	23.69%†
	$N = 40$	64.14%†	65.50%†
	$N = N_{max}$	84.24%†	86.00%†
KLD	$N = 10$	72.35%†	75.34%†
	$N = 40$	90.25%	91.03%
	$N = N_{max}$	**91.35%**	**91.73%**

Table 3: Effect of variation in N

† Significant performance difference ($\alpha = 1\%$, two-sided sign test)

5 Conclusion

Looking at the above results we can conclude that for Gujarati newspaper articles K.L.D. based authorship attribution with proper parameter selection is comparable to the current state of art Z-score based method when sufficient number of articles are available as a training set. But when the number of training examples are less then K.L.D. based method outperforms the Z-score based method. This might be because by normalising each of the terms' frequency, Z_{score} effectively considers each term to be of same importance. This might not be true as the distribution of terms that occur in most of the documents should ideally be a better signature as compared to the terms that occur in only a few documents of the author. $K.L.D.$ inherently takes into account the occurrence frequency by weighting each term with the probability of its occurrence and hence performs better.

Acknowledgement

This research is supported by part by the *Cross Lingual Information Access* project funded by *D.I.T., Government of India.*

References

M Bagavandas and G Manimannan. 2008. Style consistency and authorship attribution: A statistical investigation*. *Journal of Quantitative Linguistics*, 15(1):100–110.

John Burrows. 2002. Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

David I Holmes. 1994. Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Frederick Mosteller and David Wallace. 1964. Inference and disputed authorship: The federalist.

Fuchun Peng, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. 2003. Language independent authorship attribution using character level language models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 267–274. Association for Computational Linguistics.

Jacques Savoy. 2012. Authorship attribution based on specific vocabulary. *ACM Transactions on Information Systems (TOIS)*, 30(2):12.

George Kingsley Zipf. 1932. Selected studies of the principle of relative frequency in language.