Authorship attribution via occupancy-problem-type indices

Lukun Zheng[a], Huiqing Zheng[b] & Chandra Kundu[c],

aDepartment of Mathematics, Western Kentucky University, Bowling Green, KY, USA.

bDepartment of Modern Languages, Western Kentucky University, Bowling Green, KY, USA.

cDepartment of Mathematics, Western Kentucky University, Bowling Green, KY, USA.

Lukun Zheng, email: lukun.zheng@wku.edu.

Authorship attribution using occupancy-problem-type indices

In this paper, we propose a new methodology for authorship attribution based on a profile of indices related to the occupancy problem, called occupancy-problem indices. The occupancy problem has a long history and is an important example in standard textbooks like Feller (1971). Balls are thrown independently at K boxes according to a probability distribution $\{p_k\} = \{p_k; 1 \leq k \leq K\}$ with the probability p_i of hitting the k-th box. This process is repeated for m times and the number H_m of nonempty boxes is recorded. Calculating the probability distribution of H_m is generally referred to as the occupancy problem. We base our methodology on function words. We establish a testing procedure by constructing a confidence band of the occupancy-problem indices using the sampling distribution of H_m. We validate our proposed methodology using several writing samples whose authorship is known. We then apply this methodology to explore the question of who wrote the 15th Oz book, which has a disputing authorship between Lyman Frank Baum (1856–1919) and his successor on the Oz series Ruth Plumly Thompson (1891–1976).

Keywords: authorship attribution; occupancy problem; confidence band.

Introduction

Given a writing sample of dispute authorship, authorship attribution is the task to identify the author based on the stylistic features shown in the writing sample. The general approach to authorship attribution is to use style characteristics extracted from the text to train a classifier. There has been many applications of authorship attribution such as plagiarism detection (Vysotska et al. 2017), cyber-crime investigation (Zheng et al., 2009), and social media forensics (Rocha et al., 2017). Many authorship attribution techniques are developed based on an index about textual features of the text. In these techniques, researchers evaluated this index in two writing samples: one with dispute authorship and the other is a corpus of the proposed author's work. Afterwards, a hypothesis test is performed to determine if the values of this index on these two writing samples are statistically different. For such techniques, a key question is which index to use. Many indices have been proposed in literature, see Grieve (2007), Stamatatos (2009), Grabchak et al (2013), Zheng and Zheng (2019) and references therein. We propose a novel approach for authorship attribution based on a profile of indices related to the occupancy problem, called the occupancy-type indices. The purpose of this paper is threefold. First, we introduce the occupancy problem and explain how it can be used for authorship attribution. Second, we establish a testing procedure by constructing a confidence band of the occupancy-problem indices using the sampling distribution of related statistics. Third, we validate our proposed methodology using several writing samples with known authorship and then apply this methodology to explore the question of who wrote the fifteenth Oz book, which has a disputing authorship between Lyman Frank Baum (1856–1919) and his successor on the Oz series Ruth Plumly Thompson (1891–1976).

In authorship attribution, we start by text cleaning followed by feature extraction and normalization. Two key components of authorship attribution techniques are the choice of stylometric features in use and the type of classifier employed. Segarra et al. (2015) use function words to build stylometric fingerprints. They consider the relational structure among function words and encode those structures as word adjacency networks (WANs) which are asymmetric networks that store information of co-appearance of two function words in the same sentence. With proper normalization, edges of these networks describe the likelihood that a particular function word is encountered in the text given that we encountered another one, which implies a Markovian model. They measure the dissimilarity between different texts in terms of the relative entropy between the associated Markov chains. Brocardo et al. (2017)

explores the use of deep belief networks for authorship verification model applicable for continuous authentication. They use Gaussian units in the visible layer to model real-valued data on the basis of a Gaussian-Bernoulli deep belief network. The lexical, syntactic, and application-specific features are explored, leading to the proposal of a method to merge a pair of features into a single one.

Feature selection is an important issue for authorship attribution. We use function words as our features in our proposal methodology. In authorship attribution, content words can be misleading, as different authors writing on similar topics may use many common content words. However, function words are words such as prepositions, conjunctions, or articles that have little lexical meaning or have ambiguous meaning and express grammatical relationships among other words within a sentence, or specify the attitude or mood of the speaker. Function words can reveal the characteristics of an author's writing style. The function words prove to be reliable features in authorship attribution, see Holmes et al. (2001), Binongo (2003), Garcia and Martin (2006), and Segarra et al. (2015)

Consider the collection of all possible function words that are available in a writing sample. Let us denote the collection by $\{v_k\} = \{v_k; 1 \leq k \leq K\}$, where v_k stands for a the k-th function word. Let $\{p_k\} = \{p_k; 1 \leq k \leq K\}$ be the probability distribution of a particular author over $\{v_k\}$. It is called the 'function word distribution' of the author. For a given author, we think of p_k as the probability that the next function word the author will use is v_k. This probability is unknown and can only be estimated by certain statistics such as the relative frequency of v_k in the author's writings. An underlying assumption for our methodology is that the function words of a writing sample forms a random sample from the collection of function words $\{v_k\}$, according to the 'function word distribution $\{p_k\}$. Even though it may seem like a strong assumption, it is implicitly made in many, if not all, authorship studies based on lexical features. For a detailed discussion of this assumption, interested readers are referred to Grabchak et al. (2013).

The above suggests that we can perform authorship attribution using the underlying function word distributions. If the function words in a writing sample with dispute authorship have relative frequencies that are close to the function word distribution of a particular author, then we may conclude that the writing sample is written by this author. However, the difficulty lies in the fact that it is impossible to have an accurate estimation for the probabilities of rare function words. For this reason, instead of using the function word distribution directly, it is usually preferred to use a measure related to the distribution. While there are many such measures, we focus on a set of measures induced by the well-known occupancy problem.

The occupancy problem and its generalizations are of traditional and recurrent interests. In classical multinomial occupancy scheme, m balls are independently thrown at K boxes according to a probability distribution $\{p_k\} = \{p_i; 1 \leq k \leq K\}$. The allocation of the m balls among the K boxes is captured by the random vector $(X_{m1}, X_{m2}, \ldots, X_{mK})$, where X_{mk} is the number of balls out of the m balls that fall in box k. Let H_m be the number of boxes which has at least one ball. That is, $H_m = \sum_{k=1}^{K} I(X_{mk} > 0)$, where $I(X_{mk} > 0$ equals 1 if $X_{mk} > 0$ and it equals zero if $X_{mk} = 0$. In other words, H_m is the number of occupied boxes. The problem of studying the distribution of H_m is generally referred to as the occupancy problem. The occupancy problem has many applications in different scientific fields. If instead of boxes one has types or species of sampling units, it becomes species sampling problems in ecology and in database query optimization, where the sampling units may be entries of a database while the species are the distinct values appearing in a database, see Chaudhuri et al. (1998). In disclosure risk assessment, the sampling units may be individuals or companies listed in a microdata file without identifying information, while the types are unique combinations of values of variables with which the individuals or companies might be identified indirectly, see Skinner, & Elliot, (2002). For a further discussion of the applications see Bunge, & Fitzpatrick, (1993), Roberts, & Tesman, (2009) and references therein. In the current study, the sampling units are function words while types are distinct function words in a given writing sample.

Given the 'function word distribution' $\{p_k\} = \{p_k; 1 \leq k \leq K\}$ of a particular author over the set $\{v_k;\ 1 \leq k \leq K\}$ of all function words used by this author, let H_m be the number of distinct function words among a random sample of m function words from the corresponding function word distribution. With this notation, H_{10} is the number of distinct function words in a random sample of 10 function words written by the corresponding author. The expectation and variance of H_m are given as follows:

$$\mu_m = E(H_m) = \sum_{k=1}^{K}(1-(1-p_k)^m), \qquad (1)$$

$$\sigma_m^2 = Var(H_m) = \mu_{2m} - \mu_m + \sum_{j \neq k}\left[\left(1-p_j-p_k\right)^m - \left(1-p_j\right)^m(1-p_k)^m\right] \qquad (2)$$

see Gnedin et al (2007).

From these equations, we can see that these values capture important properties of the function word distribution and hence the writing characteristics of a given author. In the next section, we show our methodology for authorship attribution. In the third section, we present a data analysis to validate our proposed methodology and then to apply this methodology to explore the question of who wrote the fifteenth Oz book. Finally, in the fourth section, we present some concluding remarks about this methodology.

Methodology

Given a writing sample, let $\{x_i;\ 1 \leq i \leq N\}$ be the sequence of function words in order, where N is the total number of function words. Let f_k be the number of times that the function word v_k appears in the sample and let $\hat{p}_k = f_k/N$ be the relative frequency. A naïve estimator of μ_m is given by the so-called 'plug-in' estimator $\hat{\mu}_m = \sum_{k=1}^{K}(1-(1-\hat{p}_k)^m)$. Unfortunately, this estimator may have quite a bit bias.

We propose an estimator of $\hat{\mu}_m$ based on the law of large numbers. The setup is as follows:

1. A writing sample of length N of a particular author is obtained. It is assumed to be a random sample from the word type distribution consisting of the probability values $\{p_k; 1 \leq k \leq K\}$.
2. The empirical distribution $\{\hat{p}_k; 1 \leq k \leq K\}$.is obtained. It is also called the resampling distribution, since we are to collect random samples from this empirical distribution.
3. A random sample of size m is drawn from the resampling distribution and the number H_m of distinct function words in the sample is obtained.
4. We repeat step 3 independently for n times and obtain n values of H_m: $H_{m1}, H_{m2}, \cdots, H_{mn}$

This setup is validated by the law of large numbers: for sufficiently large N, the resampling distribution $\{\hat{p}_k; 1 \leq k \leq K\}$ is approximately the same as the underlying true distribution $\{p_k; 1 \leq k \leq K\}$. Since $H_{m1}, H_{m2}, \cdots, H_{mn}$ are independent and identically distributed, we have, due to central limit theorem,

$$\frac{\overline{H_m}-\mu_m}{S_m/\sqrt{n}} \xrightarrow{L} N(0,1) \quad as\ n \to \infty, \tag{3}$$

where $\overline{H_m}$and S_m are the sample mean and sample standard deviation of $H_{m1}, H_{m2}, \cdots, H_{mn}$.

We now show how to use these results for authorship attribution. Assume that we have two writing samples, Sample 1 and Sample 2, that were written independently, and we want to check if they were written by the same author. Assume that the first sample comprises N_1 function words and that the second comprises N_2 function words. In addition, we assume that the first sample contains K_1 different types of function words and the second sample contains K_2 different types of function words. We begin by choosing r, the length of our profile, which must satisfy $r \leq \min\,(N_1, N_2)$ where $\min\,(N_1, N_2)$ is the minimum of N_1 and N_2. Let $H_{m1}^{(1)}, H_{m2}^{(1)}, \dots, H_{mn}^{(1)}$ be the observed numbers of distinct function words in samples of size m from the resampling distribution based on Sample 1 and let $H_{m1}^{(2)}, H_{m2}^{(2)}, \dots, H_{mn}^{(2)}$ be the observed numbers of distinct function words in samples of size m from the resampling distribution based on Sample 2, for $m = 1,2, \dots, r$. Since these values are based on different random samples, there will be some differences among them, even if the two writing samples were indeed written by the same author. For this reason, we need to develop a method to check if these differences are statistically significant. Here we propose constructing a confidence band for the differences between $\mu_m^{(1)}$ and $\mu_m^{(2)}$. If zero is always in the band, the differences are not significant and there is no sufficient evidence to conclude that the samples are written by different authors. If zero is generally outside of band, then the differences are significant and we believe that the samples are written by different authors. If zero is sometimes in the band and sometimes not, this gives partial evidence that the samples are written by different authors.

To construct the confidence band, we start with constructing a confidence interval for each m. From equation (3) it follows that an asymptotic $100(1-\alpha)\%$ confidence interval for the difference between $\mu_m^{(1)}$ and $\mu_m^{(2)}$ is given by

$$\overline{H_{,m}^{(1)}} - \overline{H_{,m}^{(2)}} \pm z_{\alpha/2}\sqrt{\frac{\left[S_m^{(1)}\right]^2}{n} + \frac{\left[S_m^{(2)}\right]^2}{n}},$$

where $\overline{H_{,m}^{(i)}}$ and $S_m^{(i)}$ are the sample mean and sample standard deviation of $H_{m1}^{(i)}, H_{m2}^{(i)}, \ldots, H_{mn}^{(i)}$, for i=1,2, and $z_{\alpha/2}$ is the critical value of the standard normal distribution satisfying $P\left(Z > z_{\alpha/2}\right) = \alpha/2$ where $Z \sim N(0,1)$ has a standard normal distribution. For each $m = 1,2,\ldots,r$, we apply this separately and calculate a pointwise confidence band. In the next section we illustrate our methodology with an example.

Data Analysis

Lyman Frank Baum (1856–1919) was regarded as "America's greatest writer of children's fantasy." "His *Wonderful Wizard of Oz* has long been the nation's best known, best loved native fairy tale" (Gardner and Nye 1957). In about two decades, Baum published a series of 14 books of Oz, which earned him the title "The Royal Historian of Oz." He passed away on May 5, 1919. After Baum's death, the publishers Reilly & Lee (formerly Reilly & Britton) had to find someone to continue writing stories about Ozma's reign. A twelve-year-old boy called Jack Snow offered to be the next Royal Historian of Oz but was politely turned down by Reilly & Lee. Instead, they found Ruth Plumly Thompson (1891–1976), an established children's writer, as Baum's successor. Thompson wrote a new book of Oz for every Christmas season from 1921-1939. Thompson had published 33 Oz books by 1939.

All these Oz books have clear authorship except the fifteenth Oz book: The Royal Book of Oz, which was published in 1921. Baum's name was on the cover, and

Thompson was acknowledged only as having “enlarged and edited” the work. However, Oz chronicler Jack Snow (1954) believed that Thompson did *not* base the story on any notes Baum left behind and *The Royal Book of Oz* was entirely her own work. Del Rey edition in 1985 credits Thompson as the author of the book. However, a Dover edition of the book in 2001 credits Baum as the author, saying that this is “an unabridged republication of the [1921] work.” Can modern authorship attribution techniques help shed light on the authorship of the fifteenth Oz book?

the	with	up	into	just
and	but	no	now	well
to	for	out	down	where
a	at	what	over	before
an	this	then	back	upon
of	these	if	or	about
in	so	there	well	after
that	those	by	which	more
it	on	who	how	why
not	from	when	here	some
as	one	ones	all	

Table 1. Fifty four function words used in the study.

In order to use our methodology to determine the authorship of the fifteenth book, we must construct suitable corpora: one corpus consisting of 7 Oz books by Baum, and the other corpus consisting of 7 Oz books by Thompson. These books and other books used below were obtained and used under the terms of the Project Gutenberg License within corresponding books or online at www.gutenberg.org. A pre-process is applied to convert the all words into lower cases and filter all the punctuation Marks, non-ASCII character, and white space. Since we took account into only the function words, we also replaced the contraction with original words. A computer program was then written to count the number of occurrences of each of the words used. Of these words, 54 function words listed in Table 1 were selected and used. Though the reasons for choosing them were briefly discussed in the introduction, people may refer to Binongo (2003) for a detailed explanation.

Before testing the authorship of the fifteen book "The Royal Book of Oz", we validate our approach by using it for authorship attribution to several books of known authorship. Our methodology has one tuning parameter: r, which is the number of occupancy-problem-statistics in the profile. This turning parameter should be strictly less than the number of words in each writing sample under study. In fact, we have found that r should be significantly smaller than this number to ensure that the asymptotic confidence bands are accurate. For all of our tests, we fix $r = 100$. The sample size in each test is fixed to be $n=1000$.

We begin with two books by Baum. One book is another Oz book "The Tin Woodman of Oz" by Baum and the other is "Sky Island", which is not in the Oz series. In Figure 1 we present our results. The plot on the left in Figure 1 shows the difference between the profile of occupancy-problem-statistics in the Baum Oz corpus and those in the book

"The Tin Woodman of Oz". The plot on the right in Figure 1 shows the difference between the profile of occupancy-problem-statistics in the Baum Oz corpus and those in the book "Sky Island". The dashed line represents the line at zero. Since zero is almost entirely in the confidence band, there is no evidence to suggest that the two samples have different authors.

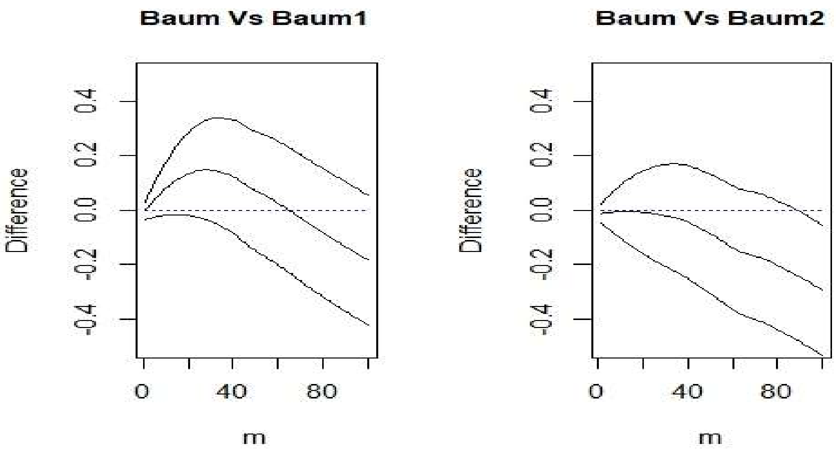

Figure 1: Methodology validation using two books by Baum. The plot on the left gives the 95% confidence band of the difference between the profiles of occupancy-problem-statistics in the Baum Oz corpus and those in the book "The Tin Woodman of Oz", denoted by "Baum1". The plot on the right in Figure 1 shows the difference between the profile of occupancy-problem-statistics in the Baum Oz corpus and those in the book "Sky Island", denoted by "Baum2".

We now validate our methodology using two books written by another contemporary author H.G. Wells. The two books are "Tales of Space and Time" and

"When the Sleeper Wakes". We compared this two books with Baum's Oz corpus and Thompson's Oz corpus. The results are given in Figure 2. The top-left plot shows the difference between the profiles of occupancy-problem-statistics in the Baum Oz corpus

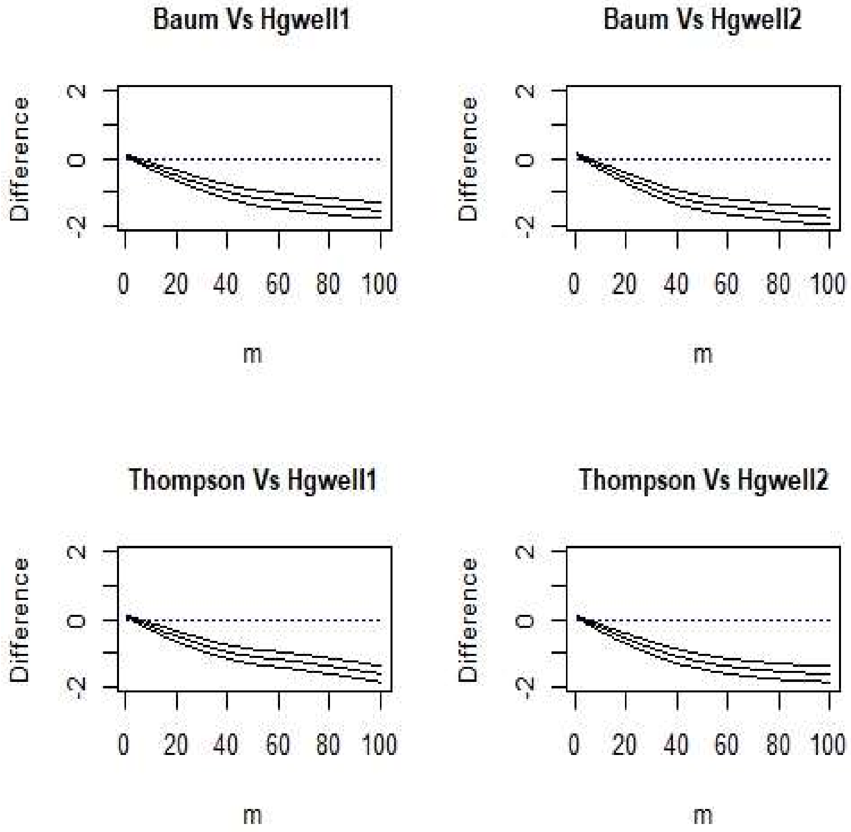

Figure 2: Validation using two books by H.G. Well. These plots give the 95% confidence bands of the difference between the profiles of occupancy-problem-statistics. The top plots are for Baum Oz corpus and Thompson Oz corpus against "Tales of Space and Time" by H.G. Well, denoted by "Hgwell1", respectively. The bottom plots are for Baum Oz corpus and Thompson Oz corpus against "When the Sleeper Wakes" by H.G. Well, denoted by "Hgwell2", respectively.

and those in the book "Tales of Space and Time". The top-right plot shows the difference between the profiles of occupancy-problem-statistics in the Baum Oz corpus and those in the book “When the Sleeper Wakes”. The bottom-left plot shows the difference between the profiles of occupancy-problem-statistics in the Thompson Oz corpus and those in the book “Tales of Space and Time”. The bottom-right plot shows the difference between the profiles of occupancy-problem-statistics in the Baum Oz corpus and those in the book “When the Sleeper Wakes”. The dashed lines represent the line at zero. Note that zero is generally outside of the confidence band and the plots clearly suggests that the authors are different. Interestingly, if we examine these plots, we may note that, technically, zero is inside of the band for a few values of m. More specifically, zero is inside the band when m is small in all plots. It shows the importance of drawing conclusion based on an entire profile instead of just one index. If we had based our conclusions on only one index, and it happened to be one that has zero inside of the entire confidence band, then we would not have enough evidence to conclude that the authors are different. However, if we base our conclusion on the entire profile, then we would.

Finally, we are ready to tackle the problem of the authorship of The Royal Book of Oz. The results are given in Figure 3. The left plot in Figure 3 shows the difference between the profiles of occupancy-problem-statistics in the Baum Oz corpus and those in the book “The Royal Book of OZ”. The right plot in Figure 3 shows the difference between the profiles of occupancy-problem-statistics in the Thompson Oz corpus and those in the book “The Royal Book of OZ”. The dashed line represents the line at zero. Note that zero is generally outside of the confidence band in left plot which provide enough evidence to support the conclusion that “The Royal Book of Oz” was written by a writer other than Baum. The fact that zero is generally inside the confidence band in Plot (b) suggests that we should credit Thompson as the author of “The Royal Book of Oz”.

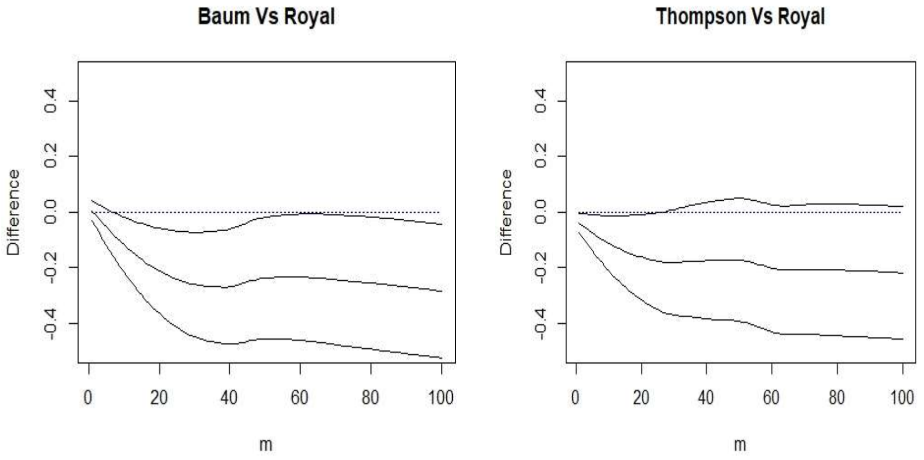

Figure 3: The 15th Oz book: The Royal Book of OZ. The left plot gives the 95% confidence band of the difference between the profiles of occupancy-problem-statistics in the Baum Oz corpus and those in the book "The Royal Book of OZ". The right plot gives the 95% confidence band of the difference between the profiles of occupancy-problem-statistics in the Thompson Oz corpus and those in the book "The Royal Book of OZ".

Conclusion

In this paper, we introduced a new methodology for authorship attribution using a profile of occupancy-problem-type indices. We applied this methodology to test if the fifteenth Oz book "The Royal Book of Oz" was written by Lyman Frank Baum (1856–1919) or by Ruth Plumly Thompson (1891–1976). We validated our methodology using several books of known authorship. However, it is important to note that our validation was only on a small scale. More extensive studies are needed to further understand the advantages and limitations of this methodology. Nevertheless, our results are sufficient to show the advantage of using an entire profile of indices instead of just one index.

This was especially evident when testing the two books “Tales of Space and Time” “When the Sleeper Wakes” by H.G. Well against the Baum Oz corpus and Thompson Oz corpus used in the experimental study.

Notes

1. All the books used in this study are no long under copyright and are downloaded from Project Gutenberg’s website: https://www.gutenberg.org/.

2. The titles of all books used to construct the Baum Oz corpus and the Thompson Oz corpus are given in Appendix.

References

Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F. (2002, March). An experiment in authorship attribution. In 6th JADT(pp. 29-37).

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance, 16(2), 9-17.

Brocardo, M. L., Traore, I., Woungang, I., & Obaidat, M. S. (2017). Authorship verification using deep belief network systems. International Journal of Communication Systems, 30(12), e3259.

Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: a review. Journal of the American Statistical Association, 88(421), 364-373.

Chaudhuri, S., Motwani, R., & Narasayya, V. (1998, June). Random sampling for histogram construction: How much is enough?. In ACM SIGMOD Record (Vol. 27, No. 2, pp. 436-447). ACM.

De Vel, O. (2000, August). Mining e-mail authorship. In Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD’2000).

Feller, W. (2008). An introduction to probability theory and its applications. John Wiley & Sons.

Garcia, A. M., & Martin, J. C. (2006). Function words in authorship attribution studies. Literary and Linguistic Computing, 22(1), 49–66.

Gardner, M. and Nye, R.B. (1957), The Wizard of Oz & Who He Was. East Lansing: Michigan State University Press.

Gnedin, A., Hansen, B., & Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. Probability surveys, 4, 146-171.

Grabchak,M., Zhang, Z.,&Zhang, D. T. (2013). Authorship attribution using entropy. Journal of Quantitative Linguistics, 20, 301–313.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, 22(3), 251–270.

Holmes, D. I., Robertson, M., & Paez, R. (2001). Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. Computers and the Humanities, 35(3), 315-331.

Roberts, F., & Tesman, B. (2009). Applied combinatorics. Chapman and Hall/CRC.

Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., ... & Stamatatos, E. (2017). Authorship attribution for social media forensics. IEEE Transactions on Information Forensics and Security, 12(1), 5-33.

Segarra, S., Eisen, M., & Ribeiro, A. (2015). Authorship attribution through function word adjacency networks. IEEE Transactions on Signal Processing, 63(20), 5464-5478.

Skinner, C. J., & Elliot, M. J. (2002). A measure of disclosure risk for microdata. Journal of the Royal Statistical Society: series B (statistical methodology), 64(4), 855-867.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556.

Vysotska, V., Burov, Y., Lytvyn, V., & Demchuk, A. (2018, August). Defining Author's Style for Plagiarism Detection in Academic Environment. In 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP) (pp. 128-133). IEEE.

Zhao, Y., & Zobel, J. (2005). Effective and scalable authorship attribution using function words. In Asia Information Retrieval Symposium (pp. 174-189). Springer, Berlin, Heidelberg.

Zheng, L. & Zheng, H (2019): Authorship Attribution via Coupon-Collector-Type Indices, Journal of Quantitative Linguistics, DOI: 10.1080/09296174.2019.1577939

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American society for information science and technology, 57(3), 378-393.

Appendix

In this appendix, we give the titles of the books used to construct the corpora in this study, together with their corresponding E book # in Project Gutenberg. The seven Oz books by L. Frank Baum used to construct the Baum's corpus are:

1. Glinda of Oz (Ebook # 961)
2. Little Wizard Stories of Oz (Ebook # 25519)
3. Ozma of Oz (Ebook # 33361)
4. The Emerald City of Oz (Ebook # 517)
5. The Lost Princess of Oz (Ebook # 24459)
6. The Patchwork Girl of Oz (Ebook # 955)
7. The Wonderful Wizard of Oz (Ebook # 55)

The seven Oz books by Ruth Plumly Thompson used to construct the Thompson Oz corpus are:

1. Ozoplaning with the Wizard of Oz- (Ebook # 55806)
2. Captain Salt in Oz (Ebook # 56073)
3. Kabumpo in Oz (Ebook # 53765)
4. The Silver Princess in Oz (Ebook # 56085)
5. Handy Mandy in Oz (Ebook # 56079)
6. The Wishing Horse of Oz (Ebook # 55851)
7. The Cowardly Lion of Oz (Ebook # 58765)