Authorship attribution via occupancy-problem-type indices

Chandra S Kundu

Advisor: Dr. Lukun Zheng

Western Kentucky University

March 5, 2020

Overview

What is Authorship Attribution?

Given a writing sample of disputed authorship, authorship attribution is the task to identify the author based on the stylistic features shown in the writing sample.

Authorship Attribution Example

Who wrote the 15^{th} book of the Oz series?

Figure: The cover of The Royal Book of Oz(1921) Source: Wikipedia

- L. Frank Baum?
  - The original author of the Oz series
  - Wrote the first 14 books of the series
  - Died in 1919.
- Ruth Plumly Thompson?
  - Wrote around 20 books of the series

Why is Authorship Attribution important?

- Plagiarism detection
- Cyber crime investigation
- Social media forensics

Approach To Authorship Attribution

- Manually research of the text by scholars - Time Consuming
- Statistical Approach
  - Step 1: developing an index based on textual features of the text
  - Step 2: Evaluating the index in two writing samples: one with disputed authorship and the corpus of the proposed authors' works
  - Step 3: Performing a hypothesis test if the values of this index on these writing samples are statistically different

Approach To Authorship Attribution

For the general statistical approach:

- Style characteristics extracted from the text - which feature to use.(Feature Selection)
- A classifier - which index to use.

Previous Work

- Segarra et al. (2015) used function words to build stylometric fingerprints using Markovian model. They measure the dissimilarity between different texts in terms of the relative entropy between the associated Markov chains.
- Brocardo et al. (2017) used lexical, syntactic, and application-specific features. They use Gaussian units in the visible layer to model real-valued data on the basis of a Gaussian-Bernoulli deep belief network.
- etc.

Feature Selection

- Function words - words such as prepositions, conjunctions, or articles that have little lexical meaning
  - can reveal the characteristics of an author's writing style
  - express grammatical relationships among other words within a sentence
  - specify the attitude or mood of the speaker.
- Content words - words such as noun, verb, adjective, adverb that have semantic content.
  - can be misleading, as different authors writing on similar topics may use many common content words
  - computationally inefficient due to their quantity

The **trees** along the **river** are **beginning** to **blossom**.

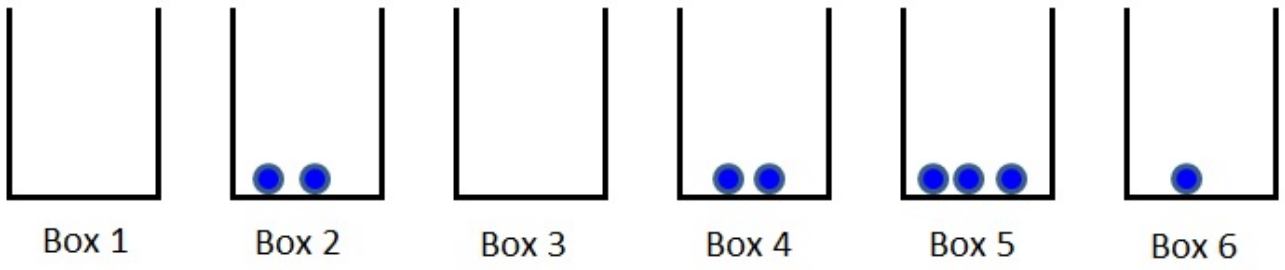
Content words are shown in bold.

Function Words Distribution

Let's consider the collection of all possible function words that are available in a writing sample.

- $\{v_k\} = \{v_k : 1 \leq k \leq K\}$, where v_k stands for a the k-th function word
- $\{p_k\} = \{p_k : 1 \leq k \leq K\}$ be the probability distribution of a particular author over $\{v_k\}$.
- 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.

Occupancy-problem-type indices



- Let, m balls are independently thrown at K boxes according to a probability distribution $\{p_k\} = \{p_i : 1 \leq k \leq K\}$.
- X_{mk} is the number of balls out of the m balls that fall in box k. So, the allocation of balls: $(X_{m1}, X_{m2}, \ldots, X_{mK})$
- Let, H_m is the number of occupied boxes. That is, $H_m = \sum_{k=1}^{K} I(X_{mk} > 0)$, where I = 1 if $X_{mk} > 0$ and otherwise 0

The study the distribution of H_m is generally referred to as the occupancy problem.

Function words and Occupancy-problem-type indices

- Given the 'function word distribution' $\{p_k\} = \{p_i : 1 \leq k \leq K\}$ of a particular author over the set $\{v_k\} = \{v_k : 1 \leq k \leq K\}$ 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.

For example, 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 (Gnedin et al (2007)):

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

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

Assumption

Distribution of function words from a random sample from the collection of function words is same as the distribution from writing samples. Even though it may seem like a strong assumption, it is implicitly made in many, if not all, authorship studies based on lexical features.

Methodology (Part 1)

1. A writing sample of a particular author is obtained and has N function words.
2. 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. The empirical distribution $\{\hat{p_k}\} = \{p_i : 1 \leq k \leq K\}$ 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 $H_m : H_{m1}, H_{m2}, \cdots, H_{mn}$

Methodology (Part 2)

- By the law of large numbers: for sufficiently large N, the resampling distribution $\{\hat{p_k}\}$ is approximately the same as the underlying true distribution $\{p_k\}$.
- Since $H_{m1}, H_{m2}, \cdots, H_{mn}$ are independent and identically distributed. From central limit theorem,

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

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

Methodology (Part 3)

- For authorship attribution, assume that we have two writing samples: Sample 1 and Sample 2.
- Choose the length of our profile, $r \leq min(N_1, N_2)$, where N_1 and N_2 are the number of function words in each sample.
- Now find
  for sample 1: $H_{m1}^{(1)}, H_{m2}^{(1)}, ..., H_{mn}^{(1)}$
  for sample 2: $H_{m1}^{(2)}, H_{m2}^{(2)}, ..., H_{mn}^{(2)}$
  for m= 1, 2, ..., r

Methodology (Part 4)

- Construct the confidence band using an asymptotic $100(1-\alpha)\%$ confidence interval for the difference between $\mu_m^{(1)}$ and $\mu_m^{(2)}$:

$$\overline{H_m^{(1)}} - \overline{H_m^{(2)}} \pm Z_{\alpha/2}\sqrt{\frac{[s_m^{(1)}]^2}{n} + \frac{[s_m^{(2)}]^2}{n}}$$

- Result from the confidence interval:
  1. If zero is always in the band: no sufficient evidence that the samples are written by different authors.
  2. If zero is generally outside of band: the samples are written by different authors.
  3. If zero is sometimes in the band and sometimes not: partial evidence that the samples are written by different authors

Data Analysis

Function words to be used

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

Figure: Fifty four function words are used in the study

To determine the authorship of the 15^{th} book of Oz series, we construct corpora: one corpus consisting of randomly chosen 7 Oz books by Baum, and the other corpus consisting of randomly chosen 7 Oz books by Thompson.

Data Analysis

Methodology validation using two books by Baum:

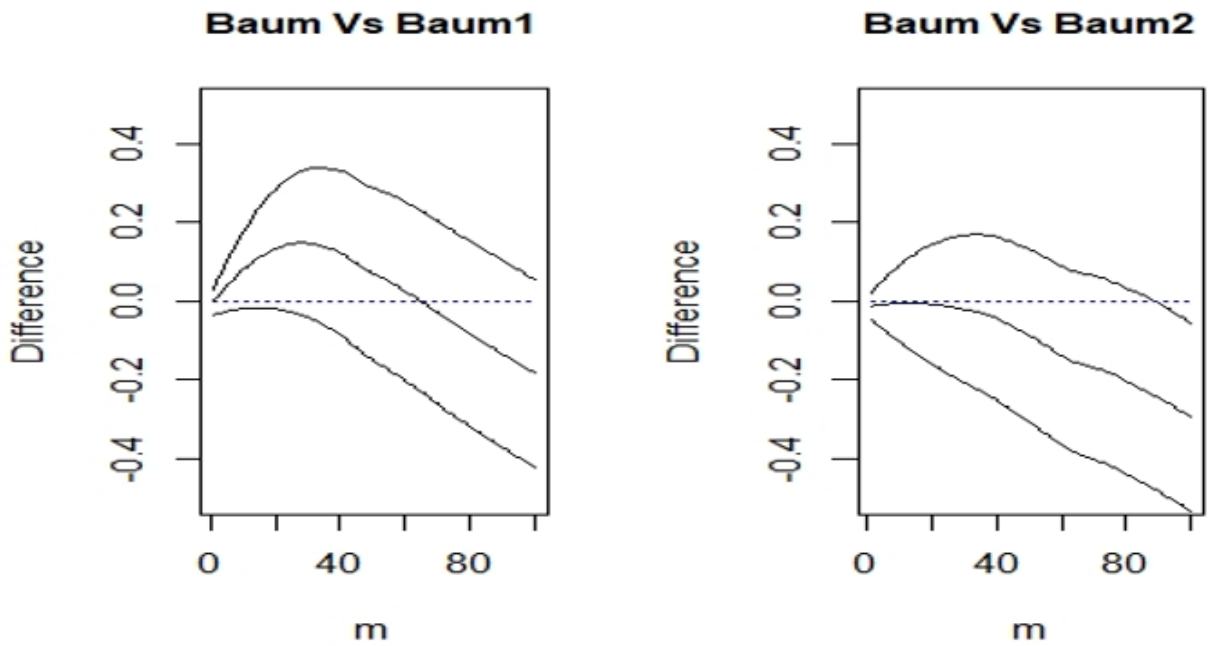
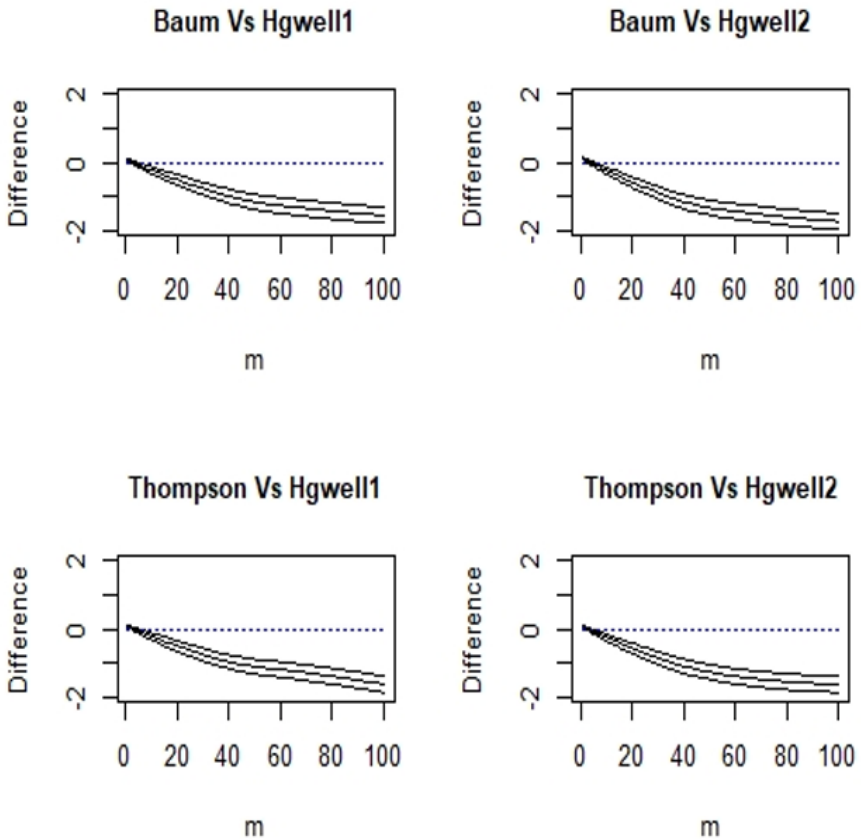


Figure: **Left:** Baum Oz corpus vs "The Tin Woodman of Oz", denoted by "Baum1". **Right:** Baum Oz corpus vs "Sky Island", denoted by "Baum2".

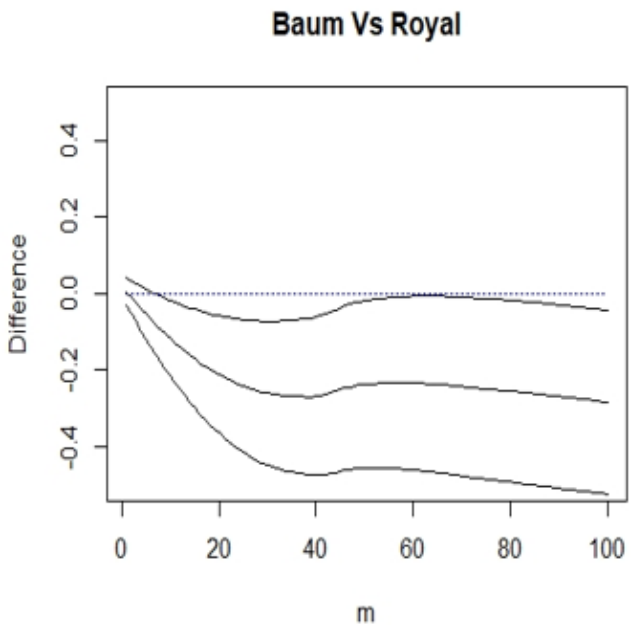
Data Analysis

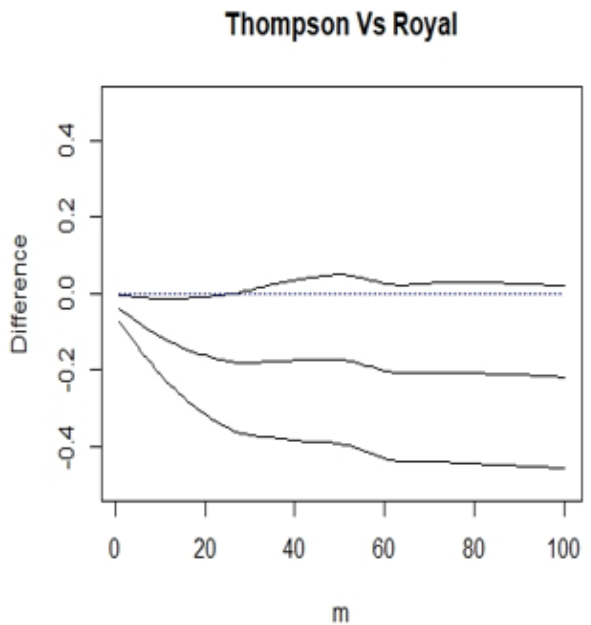
Methodology validation using two books by H.G. Well("Tales of Space and Time" as "Hgwell1" and "When the Sleeper Wakes" as "Hgwell2") vs Baum and Thompson Corpus:



Data Analysis

Who is the author of "The Royal Book of Oz", the 15^{th} book of Oz series?





Conclusion

- A new methodology for authorship attribution using a profile of occupancy-problem-type indices is introduced.
- This methodology has been validated using several books of know authorship and then applied to test if the 15^{th} book Oz series was written by
- Our validation was only on a small scale. More extensive studies are needed to further understand the advantages and limitations of this methodology.

Question?

The End