Explaining Delta
How do distance measures for authorship attribution work?

Stefan Evert, Thomas Proisl
Friedrich-Alexander-Universität Erlangen-Nürnberg

Christof Schöch, Fotis Jannidis, Steffen Pielström, Thorsten Vitt
Julius-Maximilians-Universität Würzburg

Lancaster, 24 July 2015
Outline

Authorship attribution

The parameters of Delta measures

Learning curves: How much data are needed?

Which words are most informative?

Outlook
Authorship attribution
(Juola 2006; Koppel et al. 2008; Stamatatos 2009)

- Identify unknown author or settle case of disputed authorship
  - Federalist papers: Hamilton vs. Madison (Mosteller and Wallace 1963)
  - Did Shakespeare really exist?
  - Robert Galbraith (The Cuckoo’s Calling) = J. K. Rowling
Authorship attribution
(Juola 2006; Koppel et al. 2008; Stamatatos 2009)

- Identify unknown author or settle case of disputed authorship
  - Federalist papers: Hamilton vs. Madison (Mosteller and Wallace 1963)
  - Did Shakespeare really exist?
  - Robert Galbraith (The Cuckoo’s Calling) = J. K. Rowling

- Which stylometric features determine the characteristic style of a literary author?
  - authorship attribution as a proxy task
  - “successful” features → particularly characteristic for author
Authorship attribution
(Juola 2006; Koppel et al. 2008; Stamatatos 2009)

- Authorship attribution as classification task
  - closed set of candidate authors for unknown text
  - training set of texts with known authorship
  - evaluation: classification accuracy

- Authorship attribution as clustering task
  - given set of unknown texts
  - group texts written by same author into cluster
  - evaluation: adjusted Rand index (ARI)

more general approach
Authorship attribution
(Juola 2006; Koppel et al. 2008; Stamatatos 2009)

- Authorship attribution as **classification** task
  - closed set of candidate authors for unknown text
  - training set of texts with known authorship
  - evaluation: classification **accuracy**

- Authorship attribution as **clustering** task
  - given set of unknown texts
  - group texts written by same author into cluster
  - evaluation: adjusted Rand index (ARI)
  - more general approach
Authorship attribution
(Juola 2006; Koppel et al. 2008; Stamatatos 2009)

- Popular approach: supervised machine learning
  - wide range of stylometric features
  - ML trained on texts with known authorship
  - feature selection & weighting

But not suitable for clustering task
- no supervised training data available
- clustering based on stylometric distance between texts (metric)
- no easy way to determine feature weights for metric

Simple Delta measure (Burrows 2002) very successful
Authorship attribution
(Juola 2006; Koppel et al. 2008; Stamatatos 2009)

- Popular approach: supervised machine learning
  - wide range of stylometric features
  - ML trained on texts with known authorship
  - feature selection & weighting

- But not suitable for clustering task
  - no supervised training data available
  - clustering based on stylometric distance between texts (metric)
  - no easy way to determine feature weights for metric
Authorship attribution
(Juola 2006; Koppel et al. 2008; Stamatatos 2009)

▶ Popular approach: supervised machine learning
  ▶ wide range of stylometric features
  ▶ ML trained on texts with known authorship
  ▶ feature selection & weighting

▶ But not suitable for clustering task
  ▶ no supervised training data available
  ▶ clustering based on stylometric distance between texts (metric)
  ▶ no easy way to determine feature weights for metric

▶ Simple Delta measure (Burrows 2002) very successful
Burrows’s Delta ($\Delta_B$)
(Burrows 2002)

- Frequencies of 100 – 5,000 most frequent words (MFW) form a “fingerprint” of an author’s style

\[
\begin{align*}
\text{f(Madding Crowd)} &= (.051, .029, .026, .027, .027, .016, .016, .014, .011, .008, .010, \ldots) \\
\text{f(Tess of the d’U.)} &= (.053, .027, .027, .028, .020, .013, .015, .014, .012, .009, .018, \ldots) \\
\text{f(Oliver Twist)} &= (.055, .032, .024, .023, .022, .012, .014, .011, .011, .013, .005, \ldots)
\end{align*}
\]
Burrows’s Delta ($\Delta_B$)  
(Burrows 2002)

- Frequencies of 100 – 5,000 most frequent words (MFW) form a “fingerprint” of an author’s style

\[
\begin{align*}
\text{f(Madding Crowd)} &= (.051, .029, .026, .027, .027, .016, .016, .014, .011, .008, .010, \ldots) \\
\text{f(Tess of the d’U.)} &= (.053, .027, .027, .028, .020, .013, .015, .014, .012, .009, .018, \ldots) \\
\text{f(Oliver Twist)} &= (.055, .032, .024, .023, .022, .012, .014, .011, .011, .013, .005, \ldots)
\end{align*}
\]
Burrows’s Delta ($\Delta_B$)  
(Burrows 2002)

- Frequencies of 100 – 5,000 most frequent words (MFW) form a “fingerprint” of an author’s style
- Standardized to z-scores to give each word equal weight

\[ z(\text{Madding Crowd}) = (0.53, -0.23, -0.32, 0.20, 1.66, -0.37, 1.04, 0.52, -0.44, -0.92, 0.03, \ldots) \]
\[ z(\text{Tess of the d’U.}) = (0.75, -0.48, -0.08, 0.51, -0.24, -0.87, 0.60, 0.41, -0.14, -0.47, 1.39, \ldots) \]
\[ z(\text{Oliver Twist}) = (1.05, 0.15, -0.71, -0.56, 0.37, -1.01, -0.06, -0.74, -0.28, 0.48, -0.94, \ldots) \]
Burrows’s Delta ($\Delta_B$)  
(Burrows 2002)

- Frequencies of 100 – 5,000 most frequent words (MFW) form a “fingerprint” of an author’s style
- Standardized to z-scores to give each word equal weight

$z(\text{Madding Crowd}) = (0.53, -0.23, -0.32, 0.20, 1.66, -0.37, 1.04, 0.52, -0.44, -0.92, 0.03, \ldots)$
$z(\text{Tess of the d’U.}) = (0.75, -0.48, -0.08, 0.51, -0.24, -0.87, 0.60, 0.41, -0.14, -0.47, 1.39, \ldots)$
$z(\text{Oliver Twist}) = (1.05, 0.15, -0.71, -0.56, 0.37, -1.01, -0.06, -0.74, -0.28, 0.48, -0.94, \ldots)$
The family of Delta measures
(Burrows 2002; Hoover 2004; Argamon 2008; Smith and Aldridge 2011)

- Burrows’s Delta = Manhattan distance (Burrows 2002)

\[ \Delta_B(D, D') = \|z(D) - z(D')\|_1 = \sum_{i=1}^{n_w} |z_i(D) - z_i(D')| \]
The family of Delta measures

(Burrows 2002; Hoover 2004; Argamon 2008; Smith and Aldridge 2011)

- Burrows’s Delta = Manhattan distance (Burrows 2002)

\[ \Delta_B(D, D’) = \| \mathbf{z}(D) - \mathbf{z}(D’) \|_1 = \sum_{i=1}^{n_w} |z_i(D) - z_i(D’)| \]

- Quadratic Delta = Euclidean distance (Argamon 2008)

\[ \Delta_Q(D, D’) = \| \mathbf{z}(D) - \mathbf{z}(D’) \|_2^2 = \sum_{i=1}^{n_w} (z_i(D) - z_i(D’))^2 \]
The family of Delta measures
(Burrows 2002; Hoover 2004; Argamon 2008; Smith and Aldridge 2011)

- **Burrows's Delta** = Manhattan distance (Burrows 2002)

  \[
  \Delta_B(D, D') = \|z(D) - z(D')\|_1 = \sum_{i=1}^{n_w} |z_i(D) - z_i(D')| 
  \]

- **Quadratic Delta** = Euclidean distance (Argamon 2008)

  \[
  \Delta_Q(D, D') = \|z(D) - z(D')\|_2^2 = \sum_{i=1}^{n_w} (z_i(D) - z_i(D'))^2 
  \]

- **Cosine Delta** = angular distance (Smith and Aldridge 2011)

  \[
  \cos \Delta_\angle(D, D') = \frac{\sum_{i=1}^{n_w} z_i(D) \cdot z_i(D')}{\|z(D)\|_2 \cdot \|z(D')\|_2} 
  \]
Experiments

„In theory, theory and practice are the same. In practice, they are not.“

- Empirical study based on data of Jannidis et al. (2015)
  - corpora of English, German and French novels
  - 75 novels per corpus: 3 novels each from 75 authors
  - early 19th C. to middle of 20th C.
- Exp. 1: Understanding the parameters of Delta measures
- Exp. 2: How much data are needed?
- Exp. 3: Supervised feature selection
Understanding the parameters of Delta

Prior work by Jannidis et al. (2015)

- Novels grouped into 25 clusters based on Delta distances
- All known Delta measures for $n_w = 100, 1000, 5000$ MFW
- Evaluation: within/between distances, cluster purity
- Best results: Cosine Delta $\Delta_\angle$ (Smith and Aldridge 2011) and the original Burrows Delta $\Delta_B$ (Burrows 2002)
- Mathematically sensible variants of Delta (Argamon 2008) are much worse than $\Delta_B$
Understanding the parameter of Delta

Prior work by Jannidis et al. (2015)

- Novels grouped into 25 clusters based on Delta distances
- All known Delta measures for \( n_w = 100, 1000, 5000 \) MFW
- Evaluation: within/between distances, cluster purity
- Best results: Cosine Delta \( \Delta_\angle \) (Smith and Aldridge 2011) and the original Burrows Delta \( \Delta_B \) (Burrows 2002)
- Mathematically sensible variants of Delta (Argamon 2008) are much worse than \( \Delta_B \)

New results (Evert et al. 2015)

- Detailed plots of \( n_w \) for \( \Delta_B, \Delta_Q \) and \( \Delta_\angle \)
- Systematic experiments with different parameters of Delta
- Evaluation: adjusted Rand index (Hubert and Arabie 1985)
Parameter: Number $n_w$ of MFW

English (z-scores)

<table>
<thead>
<tr>
<th># features</th>
<th>adjusted Rand index (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burrows Delta</td>
<td></td>
</tr>
<tr>
<td>Quadratic Delta</td>
<td></td>
</tr>
<tr>
<td>Cosine Delta</td>
<td></td>
</tr>
</tbody>
</table>

(following slides will show results for English corpus only)
Parameter: Number $n_w$ of MFW

(following slides will show results for English corpus only)
Parameter: Number $n_w$ of MFW

(following slides will show results for English corpus only)
Parameter: Standardization of relative frequencies

relative frequencies (unscaled)
Parameter: Standardization of relative frequencies

z-scores (standardized)
Parameter: Standardization of relative frequencies

$L_1$-scaling $\rightarrow$ worse
Parameter: Standardization of relative frequencies

German

association measure: Mutual Information \( \rightarrow \) much worse
Parameter: Normalization of vector length

English (z-scores)

# features
adjusted Rand index (%) Burrows Delta Quadratic Delta Cosine Delta

Euclidean distance (QD) + normalization = Cosine Delta
Parameter: Normalization of vector length

Euclidean distance (QD) + normalization = Cosine Delta
Parameter: Normalization of vector length

Normalization has crucial effect on clustering quality!
Why is vector normalization so important?

Feature vector $\mathbf{z}(D) =$ stylistic “fingerprint” of author

Our conjecture: pattern of positive/negative deviations from norm reflects individual stylistic profile of an author

Vector length $=$ degree to which individual style is expressed
Why is vector normalization so important?
Why is vector normalization so important?
Why is vector normalization so important?
Why is vector normalization so important?
Why is vector normalization so important?

![Graph showing the relationship between mean positive and mean negative scores for various authors in English (L2, 1000 mfw). The graph includes a legend with symbols for different authors such as barclay, blackmore, braddon, burnett, cbronte, chesterton, collins, corelli, dickens, doyle, eliot, forster, thackeray, trollope, and ward. The axes are labeled mean positive score and mean negative score.]
Why is vector normalization so important?
Why is vector normalization so important?
Why is vector normalization so important?
Learning curves

- All experiments carried out on complete novels so far
Learning curves

- All experiments carried out on complete novels so far
- Does authorship attribution also work for shorter texts?
  - Experiments with Cosine Delta $\Delta_{\angle}$ and $n_w = 2000$ MFW
Learning curves: Clustering task

ARI in corpus EN, (50:110000:50), 2000 mfw, Cosine Delta

all texts shortened to specified number of word tokens
Learning curves: Classification task

Accuracy of attribution in corpus EN

one text shortened, other texts used as training data
Finding the key words: recursive feature elimination

- Greedy algorithm for selection of an optimal set of features

Procedure:
- train linear support vector machine (SVM)
- based on \([0, 1]\)-scaled relative frequencies (not on z-scores)
- discard \(k\) features with lowest SVM weights

Iterative reduction of feature set
1. all recurrent words (\(df > 1\))
2. down to \(n_w = 50,000\) (\(k = 10,000\))
3. down to \(n_w = 5,000\) (\(k = 1,000\))
4. down to \(n_w = 500\) (\(k = 100\))
5. find minimal feature set by cross-validation (\(k = 1\))
Automatically selected features

Document frequencies ($df$) of selected features

English ($n_w = 233$)  German ($n_w = 240$)  French ($n_w = 370$)

▶ Many function words, but also content words (→ overtraining)
Automatically selected features

Document frequencies ($df$) of selected features

- English ($n_w = 233$)
- German ($n_w = 240$)
- French ($n_w = 370$)

- Many function words, but also content words (⇒ overtraining)
- Some text artefacts: Roman numerals (XL, XXXVII) in novels with many chapters, graphemic variation (e.g. DE *gibt* / *giebt*)
Automatically selected features

Document frequencies ($df$) of selected features

- Many function words, but also content words (overtraining)
- Some text artefacts: Roman numerals (XL, XXXVII) in novels with many chapters, graphemic variation (e.g. DE gibt / giebt)
- Key words for English novels: with, so, t, But, And, upon, don, head, Then, looking, almost, indeed, nor, London, feel, cannot, . . ., XXXVII ($df = 34$), XLI ($df = 29$), XLIII ($df = 26$), hereabout ($df = 11$), vilest ($df = 15$), contours ($df = 9$), Ecod ($df = 4$)
Validation

- Validation of German features on unseen test sets

  - Test set A: classification
    - 71 unseen novels from 19 of the 25 authors
    - Unbalanced, with singleton authors
    - Maximum Entropy classifier trained on German corpus
    - Result: 97% classification accuracy

  - Test set B: clustering
    - 155 unseen novels from 34 authors (6 seen, 28 unseen)
    - Clustering based on Cosine Delta into 34 groups
    - Result: features ARI 240 selected 87%, 2000 MFW 83%
Validation

- Validation of German features on unseen test sets

- Test set A: classification
  - 71 unseen novels from 19 of the 25 authors
  - unbalanced, with singleton authors
  - Maximum Entropy classifier trained on German corpus
  - result: 97% classification accuracy
Validation

- Validation of German features on unseen test sets

- Test set A: classification
  - 71 unseen novels from 19 of the 25 authors
  - unbalanced, with singleton authors
  - Maximum Entropy classifier trained on German corpus
  - result: 97% classification accuracy

- Test set B: clustering
  - 155 unseen novels from 34 authors (6 seen, 28 unseen)
  - clustering based on Cosine Delta into 34 groups
  - result:
    | features     | ARI |
    |--------------|-----|
    | 240 selected | 87% |
    | 2000 MFW     | 83% |
Conclusion & outlook

Further reading
Evert, Stefan; Proisl, Thomas; Jannidis, Fotis; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten (2015). Towards a better understanding of Burrows’s Delta in literary authorship attribution. In Proceedings of the Fourth Workshop on Computational Linguistics for Literature, Denver, CO.

Next steps
- Consistency: do fragments of the same text cluster?
- Performance on selected parts of speech (e.g. function words)
- Pre-processing: lemmatization, stem + suffix, ...

Interpretation of Delta
- Identify features with largest contribution to $\Delta$ clustering
- Delta measures based on general stylometric features


Evert, Stefan; Proisl, Thomas; Jannidis, Fotis; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten (2015). Towards a better understanding of Burrows’s Delta in literary authorship attribution. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver, CO.


