Authorship Attribution: Using Rich Linguistic Features when Training Data is Scarce

Ludovic Tanguy, Franck Sajous, Basilio Calderone and Nabil Hathout

CLLE-ERSS: CNRS & University of Toulouse, France

PAN 2012 – Authorship Attribution - CLEF
Overview

- **General method for all subtasks**
  - Maximum Entropy classifier (csvLearner)
  - Substantial effort in feature engineering
    - *Many linguistically rich features*
  - No feature selection
  - Whole texts as items (no splitting)

- **Four runs were submitted:**
  - Run 1 (CLLE-ERSS1): char. trigrams + all linguistic features
  - Run 2 (CLLE-ERSS2): character trigrams only
  - Run 3 (CLLE-ERSS3): bag of words (lemma frequencies)
  - Run 4 (CLLE-ERSS4): a selection of 60 synthetic features
All training and test texts were:
- Normalised for encoding
- De-hyphenised (based on a lexicon)
- POS-tagged and parsed (Stanford CoreNLP)

No split?
- Using splits of the same few texts is misleading (textual cohesion)
- No cross-validation data available...
List of features (1)

- **Contracted forms**
  - Average ratio of frequencies (« do not » vs « don’t », etc.)

- **Phrasal verbs**
  - Frequency of all verb-prepositions pairs (« put on », etc.)

- **Lexical genericity and ambiguity**
  - Average depth in WordNet
  - Average number of synsets per word

- **Frequency of POS trigrams**

- **Syntactic dependencies**
  - Frequency of all word-relation-word triples (« cat – subj – eat »)

- **Syntactic complexity**
  - Average depth of syntactic parse trees
  - Average length of syntactic links
List of features (2)

- **Lexical cohesion**
  - Density of semantically-similar word pairs
    - *(according to Distributional Memory database)*

- **Morphological complexity**
  - Frequency of suffixed words

- **Lexical absolute frequency**
  - Repartition of words according to Nation’s wordlists

- **Punctuation and case**
  - Frequency of punctuation marks
  - Frequency of uppercased words

- **Direct speech**
  - Ratio of sentences between quotes

- **First person narrative**
  - Relative frequency of « I » (per verb, outside quotes)
Closed-class tasks (A,C,I)
- Choose the author with highest probability

Open-class tasks (B,D,J)
- Author is « unknown » if
  \[ \max(p) < \text{mean}(p) + 1.25 \times \text{st.dev}(p) \]

Results:
- Overall:
  - All rich+3char > synthetic rich > lemmas > 3char
- Results:
  - Good for A, I and J
  - Average for B
  - Bad for C and D
## Posthoc analysis

### Lesion studies on test data for tasks A and C

- Measuring accuracy with different combinations of features
- Average accuracy gain when adding each subset

<table>
<thead>
<tr>
<th>Feature Subset</th>
<th>Gain for task A</th>
<th>Gain for task C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punctuation &amp; case</td>
<td>+0.204</td>
<td>-0.040</td>
</tr>
<tr>
<td>Suffix frequency</td>
<td>+0.097</td>
<td>+0.009</td>
</tr>
<tr>
<td>Absolute lexical frequency</td>
<td>+0.030</td>
<td>-0.003</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>+0.015</td>
<td>+0.006</td>
</tr>
<tr>
<td>Ambiguity/genericity</td>
<td>+0.012</td>
<td>+0.008</td>
</tr>
<tr>
<td>Lexical cohesion</td>
<td>+0.002</td>
<td>-0.000</td>
</tr>
<tr>
<td>Phrasal verbs (synthetic)</td>
<td>-0.000</td>
<td>+0.022</td>
</tr>
<tr>
<td>Morphological complexity</td>
<td>-0.005</td>
<td>-0.002</td>
</tr>
<tr>
<td>Phrasal verbs (detail)</td>
<td>-0.006</td>
<td>-0.006</td>
</tr>
<tr>
<td>Constructions</td>
<td>-0.014</td>
<td>+0.018</td>
</tr>
<tr>
<td>First/third person narrative</td>
<td>-0.027</td>
<td>-0.026</td>
</tr>
<tr>
<td>POS trigrams</td>
<td>-0.028</td>
<td>+0.045</td>
</tr>
<tr>
<td>Char. trigrams</td>
<td>-0.034</td>
<td>+0.206</td>
</tr>
<tr>
<td>Syntactic dependencies</td>
<td>-0.059</td>
<td>+0.089</td>
</tr>
</tbody>
</table>

$r = -0.48$
Author clustering / intrusion tasks

- Using MaxEnt as an *unsupervised* classifier
  - Method proposed by DePauw and Wagacha, 2008

- Principles:
  - Training: all paragraphs as training items
    - *Class value = paragraph ID*
  - Reclassifying: every paragraph processed by the trained classifier
    - *Result = square matrix of probabilities (Mp)*
    - *Distance matrix between paragraphs: Md = -log(Mp)*
  - Clustering: regroup similar paragraphs
    - *Hierarchical ascending clustering on Md*
  - Result: highest level clusters
Task F, Text 4, Run CLLE-ERSS1 (correct guess)
Conclusions

- Average results for traditional tasks, quite disappointing
- Good results for paragraph intrusions
- Overall, rich features are once more proven to be an improvement over character trigrams
- There’s still room for improvement with feature selection
  - Feature efficiency varies greatly across tasks and authors
  - Very small linguistic feature subsets can be sufficient