Authorship attribution of e-mail as a multi-class task

Notebook for PAN at CLEF 2011
Amsterdam, The Netherlands
21 September 2011

Kim Luyckx
CLiPS Computational Linguistics Group
University of Antwerp, Belgium
Authorship attribution

- Delicate balance between
  - Discriminative features & approach
  - Scalability: sensitivity to differences in author set size, data size, text length

- Text categorization approach
  - 1. features 2. discriminative learning
  - Common in the field

- Often binary SVM classifiers: one-vs.-all or one-vs.-one
Writing style

Assumptions
- identity interacts with writing style
- aspects you are unconscious of
- analysis of writing style allows us to identify the author

Identity = mix of age, gender, personality, education level, ideology,...
Data set specifics

- SMALL and LARGE authorship identification scenarios
- Challenging materials (Enron E-mail Corpus)
  - Quite a large group of suspects (26 and 72, resp.)
  - Short texts (+/- 60 words/e-mail)
  - Skewed class distributions (10,000 words in 200 e-mails vs. 500 words in 10 e-mails)
  - Small-world data set but a lot of internal variation (meetings, financial information etc.)
Approach

- Pre-processing
  - Tokenization
  - Removed everything between `<omni> </omni>` tags
    - Lost training data for 2 authors in both scenarios

- Text categorization approach
  - Extract features & determine the most relevant ones
  - SVMs to build a model & test it on test data
Features

- CHR $n$-grams
- $n$-grams of LEX items
- **DISC:** however, nevertheless, on the contrary
- **MOD:** can, could, would, shall
- Ranking & selection
  - Chi-square for feature relevance ranking
  - Restricted to top-1000
SVM *multiclass*

- Open-source
- Model all classes simultaneously, instead of one by one
- \( C \) ‘soft margin parameter’
  - High \( C \) ~ hard-margin classification
  - Low \( C \) introduces a lot of training errors
Development results

- Without parameter tuning $C=5,000$
- Tuning of $C$ yielded no significant difference in results

<table>
<thead>
<tr>
<th>SMALL</th>
<th></th>
<th>LARGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Macro $F_1$</td>
<td>Micro $F_1$</td>
</tr>
<tr>
<td><strong>CHR3</strong></td>
<td>37.1</td>
<td>59.4</td>
</tr>
<tr>
<td><strong>LEX1</strong></td>
<td>33.1</td>
<td>54.9</td>
</tr>
<tr>
<td><strong>DISC</strong></td>
<td>4.5</td>
<td>8.6</td>
</tr>
<tr>
<td><strong>MOD</strong></td>
<td>2.0</td>
<td>6.5</td>
</tr>
<tr>
<td><strong>CHR-var</strong></td>
<td>26.9</td>
<td>49.7</td>
</tr>
<tr>
<td><strong>LEX-var</strong></td>
<td>34.0</td>
<td>57.3</td>
</tr>
<tr>
<td><strong>CHR+LEX</strong></td>
<td>31.4</td>
<td>54.1</td>
</tr>
<tr>
<td><strong>CHR3</strong></td>
<td>27.3</td>
<td>40.6</td>
</tr>
<tr>
<td><strong>LEX1</strong></td>
<td>28.8</td>
<td>42.2</td>
</tr>
<tr>
<td><strong>DISC</strong></td>
<td>1.7</td>
<td>3.4</td>
</tr>
<tr>
<td><strong>MOD</strong></td>
<td>1.7</td>
<td>4.4</td>
</tr>
<tr>
<td><strong>CHR-var</strong></td>
<td>22.0</td>
<td>35.6</td>
</tr>
<tr>
<td><strong>LEX-var</strong></td>
<td>31.2</td>
<td>46.1</td>
</tr>
<tr>
<td><strong>CHR+LEX</strong></td>
<td>24.5</td>
<td>38.2</td>
</tr>
</tbody>
</table>
Test results

**Expectations**
- CHR3 > LEX-var in SMALL
- LEX-var > LEX1 in Large

<table>
<thead>
<tr>
<th></th>
<th>SMALL</th>
<th></th>
<th>LARGE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Macro F₁</td>
<td>Micro F₁</td>
<td>Macro F₁</td>
<td>Micro F₁</td>
</tr>
<tr>
<td>CHR3 (9/17)</td>
<td>34.3</td>
<td>62.0</td>
<td>LEX1 (7/18)</td>
<td>34.0</td>
</tr>
<tr>
<td>LEX-var (6/17)</td>
<td>37.1</td>
<td>64.2</td>
<td>LEX-var (9/18)</td>
<td>34.2</td>
</tr>
<tr>
<td>WINNERS</td>
<td>47.5</td>
<td>71.7</td>
<td>WINNERS</td>
<td>52.0</td>
</tr>
</tbody>
</table>

Which features are in LEX-var?

- Dates, locations
- Expressions of politeness (*thanks, regards, you soon*)
- E-mail specifics (*attached is*)
- Pronouns
- Argumentation (*for he*)
- Company names (*Reliant, Dominion, Enpower*)
- Domain-specific words (*pipeline*)
Conclusions

- What is our ceiling?
  - What is humanly possible?
  - What is reasonably possible?

- Is it realistic to think we will get an answer?

- Severe lack of theory in the field
  - What is authorial style?
  - What do character n-grams bring us?
In reality, no one knows what writing style is independent of the genre, register, topic?
- can you recognize the author of a letter in a newspaper article?
- independent of
  - the author’s maturity in writing?
  - familiarity with the topic?
  - his/her mood?

... consequences for validity of approaches suggested!
Contact

- kim.luyckx@ua.ac.be
- http://www.clips.ua.ac.be/~kim