GLAD: Groningen Lightweight Authorship Detection

PAN, Authorship verification, 2015

Manuela Hürlinmann, Benno Weck, Esther van den Berg, Simon Šuster, Malvina Nissim
The challenge

given: a set of Known documents written by the same Author A_K,
given: one Unknown document written by an unknown Author A_U,
task: determine whether \( A_U = A_K \)
How can we recognise different authors?

Gentle Tony

Fat Vinny

The Weasel
How can we recognise different authors?

- Unusual word choice?
- Shorter sentences?
- More complex grammar?
How can we recognise different authors?

individual_vector(feat1, feat2, ...)

individual_vector(feat1, feat2, ...)

individual_vector(feat1, feat2, ...)

Gentle Tony

Fat Vinny

The Weasel
How can we then differentiate between authors?
How can we then differentiate between authors?

- Different word choice?
- Different sentence length?
- Different grammar?
How can we then differentiate between authors?

\texttt{similarity\_vector(\textit{feat1}, \textit{feat2}, \ldots)}
Our approach

• machine learning approach training on PAN (2015) data
• using SVM to do two-class classification task
• a set of features
• feature ablation studies to tune the system to each different language
The core aim

- A lightweight system!
The aim

Input in any language
The aim

Input in any language

Features should be easy to extract

Training instance

Training instance

Training instance

Model
The aim

Input in any language

Features should be easy to extract

Training & Testing time should be fast

prediction

model

training instance training instance training instance
Our features

- Entropy of known
- Entropy of unknown
- Joint entropy
- N-gram overlap
- Morpho-syntactic similarity
- Sentence length of known
- Sentence length of unknown
- Visual features
- Compression dissimilarity
- Token cosine similarity
Our features

similarity_vector(entropy_of_known, visual_features, ...)

- entropy of known
- entropy of unknown
- joint entropy
- n-gram overlap
- morphosyntactic similarity
- sentence length of known
- sentence length of unknown
- visual features
- compression dissimilarity
- token cosine similarity
Our features

To determine relevance: grouping
Our features

- **Individual**
  - Vector\_K(feat1,feat2)
  - Vector\_U(feat1,feat2)

- **Joint**
  - Vector\_Joint(feat1,feat2)

- **Features**
  - Entropy of known
  - Entropy of unknown
  - Sentence length of known
  - Sentence length of unknown
  - Joint entropy
  - N-gram overlap
  - Morpho-syntactic similarity
  - Visual features
  - Compression dissimilarity
  - Token cosine similarity
Comparing features

All Individual
- Sentence length of known
- Sentence length of unknown
- Entropy of known
- Entropy of unknown

Visual features + Compression dissimilarity

All Joint
- n-gram overlap
- Token cosine similarity
- Joint entropy
- Morpho-syntactic similarity
- Compression dissimilarity
- Visual feature

All Joint
- Visual features + n-gram overlap + Token cosine similarity
Comparing features

Results of ablation & single-feature experiments:
Helpful features
Side note: Visual features

- Punctuation
- Line ending
- Letter case
- Ling length
- Block size
Side note: Visual features

- Punctuation
- Line ending
- Letter case
- Ling length
- Block size

Con

- Not a characteristic of the author
- Not a linguistic feature
“Pa-pa, pa-pa, pa-pa!
Here, stop her. She’ll fall down.
Here, turn around. Walk this way.

Ma-ma, ma-ma, ma-ma;

Oh, I think you are a darling.

Mer-ry Christ-mas! Mer-ry Christmas.”
Comparing features

Results of ablation & single-feature experiments:
Harmful features
Comparing features

Results of ablation & single-feature experiments:
Features that are harmful, helpful, or helpful-depending-on-the-language
Comparing features

Results of ablation & single-feature experiments:
Features that are harmful, helpful, or helpful-depending-on-the-language
Comparing features

Results of ablation & single-feature experiments:

Differences are subtle
Comparing features

Results of ablation & single-feature experiments:
Differences are subtle
Resulting groups

Combo 1
- n-gram overlap
- visual features

Combo 2
- n-gram overlap
- visual features
- token cosine similarity

Combo 3
- n-gram overlap
- visual features
- token cosine similarity
- joint entropy

Combo 4
all features (excl. morpho-syntactic features)
## Results

<table>
<thead>
<tr>
<th>Language</th>
<th>Training</th>
<th>Test</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch (full set)</td>
<td>.55</td>
<td>.62</td>
<td>3</td>
</tr>
<tr>
<td>English (full set)</td>
<td>.56</td>
<td>.41</td>
<td>7</td>
</tr>
<tr>
<td>Greek (combo2)</td>
<td>.54</td>
<td>.60</td>
<td>4</td>
</tr>
<tr>
<td>Spanish (full set)</td>
<td>.90</td>
<td>.54</td>
<td>5</td>
</tr>
</tbody>
</table>
Results

- Simple similarity features work

<table>
<thead>
<tr>
<th>Language</th>
<th>Training</th>
<th>Test</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch (full set)</td>
<td>.55</td>
<td>.62</td>
<td>3</td>
</tr>
<tr>
<td>English (full set)</td>
<td>.56</td>
<td>.41</td>
<td>7</td>
</tr>
<tr>
<td>Greek (combo2)</td>
<td>.54</td>
<td>.60</td>
<td>4</td>
</tr>
<tr>
<td>Spanish (full set)</td>
<td>.90</td>
<td>.54</td>
<td>5</td>
</tr>
</tbody>
</table>
• Simple similarity features work in unison
Results

- Simple similarity features work in unison independent of language (except Greek)

<table>
<thead>
<tr>
<th>Language</th>
<th>Training</th>
<th>Test</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch (full set)</td>
<td>.55</td>
<td>.62</td>
<td>3</td>
</tr>
<tr>
<td>English (full set)</td>
<td>.56</td>
<td>.41</td>
<td>7</td>
</tr>
<tr>
<td>Greek (combo2)</td>
<td>.54</td>
<td>.60</td>
<td>4</td>
</tr>
<tr>
<td>Spanish (full set)</td>
<td>.90</td>
<td>.54</td>
<td>5</td>
</tr>
</tbody>
</table>
Results

- Simple similarity features work in unison independent of language (except greek)
- System works fast (runtime av. 1 minute)

<table>
<thead>
<tr>
<th>Language</th>
<th>Training</th>
<th>Test</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch (full set)</td>
<td>.55</td>
<td>.62</td>
<td>3</td>
</tr>
<tr>
<td>English (full set)</td>
<td>.56</td>
<td>.41</td>
<td>7</td>
</tr>
<tr>
<td>Greek (combo2)</td>
<td>.54</td>
<td>.60</td>
<td>4</td>
</tr>
<tr>
<td>Spanish (full set)</td>
<td>.90</td>
<td>.54</td>
<td>5</td>
</tr>
</tbody>
</table>
Final conclusion

GLAD

… is a light and fast language-independent system

… allows language adaptation done via feature selection

… involves innovative visual features which appear useful (especially for English data) and could be investigated further