Vote/Veto Classification, Ensemble Clustering and Sequence Classification for Author Identification

Roman Kern[1,2]
Stefan Klampfl[2] Mario Zechner[2]

[1] Knowledge Management Institute - Graz University of Technology
[2] Know-Center

rkern@tugraz.at
{rkern, sklampfl, mzechner}@know-center.at

PAN Workshop @ CLEF 2012 / 2012-09-20

Authorship Attribution - Approach

Vote/Veto Classification

- Same as last year
  $\Rightarrow$ Compare data-sets

- Three different feature-set sets
  $\Rightarrow$ Compare influence of uni-grams features vs. stylometric features

Authorship Attribution - Classification

Classification Algorithm

- Combine feature-spaces via individual base classifiers
- Based on performance in training phase
- In classification phase combine results

Base Feature Spaces

- Basic statistics, token statistics, grammar statistics
- Stop-word terms, slang terms, pronoun terms
- Intro-outro terms, bigram terms, unigram terms, terms

Feature Space Combinations

- Terms
- Stylometric
- Statistics

Authorship Attribution - Data-Sets

Basic Statistics

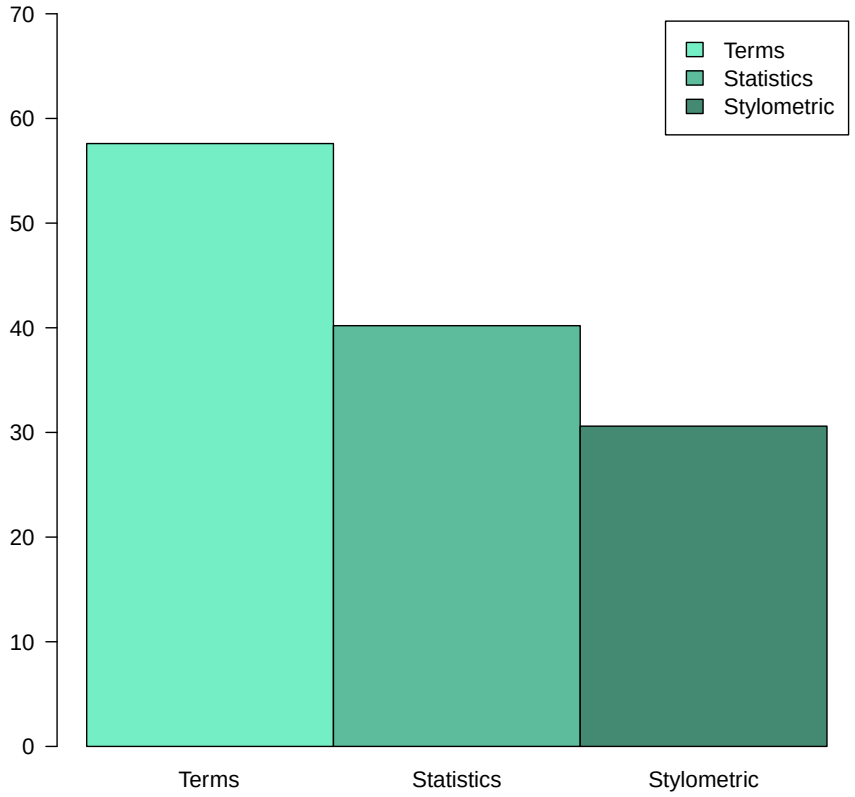
	PAN 2011	PAN 2012
1	Paragraph to lines ratio	Number of characters
2	Text to lines ratio	Number of words
3	Number of lines	Number of lines
4	Empty lines ratio	Number of stop-words
5	Number of paragraphs	Number of tokens

Token Statistics

	PAN 2011	PAN 2012
1	Likelihood of proper nouns	Number of tokens
2	Number of tokens	Likelihood of proper nouns
3	Average token length	Average verb length
4	Likelihood of infrequent word groups	Average token length
5	Likelihood of tokens of length 9	Likelihood of pronouns

Authorship Attribution - Feature Types

Comparison of configurations



Authorship Clustering - Approach

Ensemble Clustering

- Multi-tier clustering
- Combine output of base clusters
- Only use stylometric features

Ensemble clustering is also known as consensus clustering or clustering aggregation

Authorship Clustering - Features

Multiple feature spaces

- Basic statistics (same as for authorship attribution)
- Stylometric features (hapax-legomena, hapax-dislegomena, yules-k, simpsons-d, brunets-w, sichels-s, honores-h, ...)
- Stem-suffixes, stop-words, pronouns
- Character 1-grams, 2-grams, 3-grams

$\Rightarrow$ Total of 7 feature spaces

Authorship Clustering - Clustering

Base clustering

- k-means clustering
- k-means++ seed selection
- Different relatedness measures for different feature spaces
  - Cosine similarity
  - Euclidean distance (after normalising the features)

Ensemble clustering

- Create a meta-space from the individual clustering solution
- In meta-space the distance between instances depends on the agreement of the clustering solutions
  - Give different base clusters different weight
- k-means clustering

Authorship Clustering - Evaluation

Ensemble clustering results

Feature Space	A vs B	C vs D	E vs F
1-grams	51.52%	53.98%	61.87%
2-grams	50.91%	54.46%	56.70%
3-grams	50.91%	51.33%	52.37%
Stop-Words & Pronouns	62.20%	50.72%	**72.91%**
Stem Suffices	**65.85%**	**63.01%**	54.61%
Stylometry	52.74%	59.76%	64.25%
Basic Statistics	57.01%	56.87%	65.22%
Ensemble	**66.10%**	**80.34%**	**78.44%**

Sexual Predator Identification - Approach

Sequence classification

- Not directly classify predators
- Classify individual messages/line in chats
- Simple features

Sexual Predator Identification - Classes

Chat message classes/labels

- normal, predator; offending; reaction, post-offending

Chat #1
1 normal
2 predator
3 normal
4 normal
5 predator
6 normal
7 predator
8 predator
9 normal

Chat #2	
1 normal	pre
2 predator	
3 normal	
4 normal	
5 offending	post
6 reaction	
7 post-offending	
8 post-offending	
9 reaction	
10 reaction	

Sexual Predator Identification - Features

Simple features

- Unigrams
- Double Metaphone
- `isInitialAuthor`, `isLastAuthor`, `isMostVerboseAuthor`, `isFewerAuthors`, `hasTermFromPrevious`

Classification algorithm

- Maximum entropy & beam search

Sexual Predator Identification - Training

localhost/annotate/

Datei auswählen output.txt

output .txt Save

```
Conversation: cc21a4030e5be0428f302d96452a2fbd
pred: hi
pred: u found more pics
pred: can u do it tonight
pred: no
pred: do it now
pred: leave meassage my cell is dead
pred: 7077187918
pred: k
pred: ok
pred: <email/>
pred: k
```

```
Conversation: 0a00d0980272a40af6e8475e88363f44
pred: i love u
```

```
Conversation: 168d5c53c212270d6cd51972abdc3511
pred: yea
pred: k
pred: huh
pred: when u call me i'll tell u what i'll be wearing
pred: wow
pred: top
pred: cool
pred: bring jacket
pred: when u leave are u going to miss me alot
pred: yeas
pred: call me tommrow
pred: what time
pred: yea
pred: 10am
pred: what time she leave
pred: work tommrow
pred: sneak
pred: yea have to sneak too
pred: aways with u
pred: casll me at 10am
pred: ok
pred: what time do u want to call
pred: no i wont
pred: lol
pred: love sneaking
pred: does anyone know where u are going
pred: not ur friend
```

Sexual Predator Identification - Results

Class	Count	Precision	Recall
normal	3,117	0.955	0.995
predator	29	0.3	0.103
offending	52	0	0
post-offending	216	0.871	0.847
reaction	275	0.959	0.764
Identify predators	2	0.667	1

The End

Thank you!

Open-source code

`https://www.knowminer.at/svn/opensource/projects/pan2012/trunk`

Corresponding Author

Roman Kern <rkern@tugraz.at>