Clustering by Authorship Within and Across Documents

E. Stamatatos, M. Tschuggnall, B. Verhoeven,
W. Daelemans, G. Specht, B. Stein, and M. Potthast

pan@webis.de

http://pan.webis.de

Supervised vs. Unsupervised Authorship Attribution

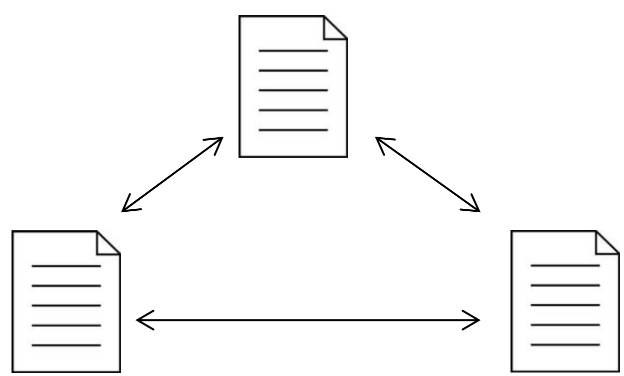
- Supervised:
  - When texts of known authorship are available
  - Labelled data
  - Closed-set and open-set attribution, Verification
- Unsupervised:
  - When authorship information either does not exist or is not reliable
  - Single-author documents -> author clustering
  - Multi-author documents -> author diarization

Lack of Reliable Authorship Information

- Examples:
  - Novels published anonymously or under an alias
  - Proclamations by different terrorist groups
  - Product reviews by different user profiles
  - …

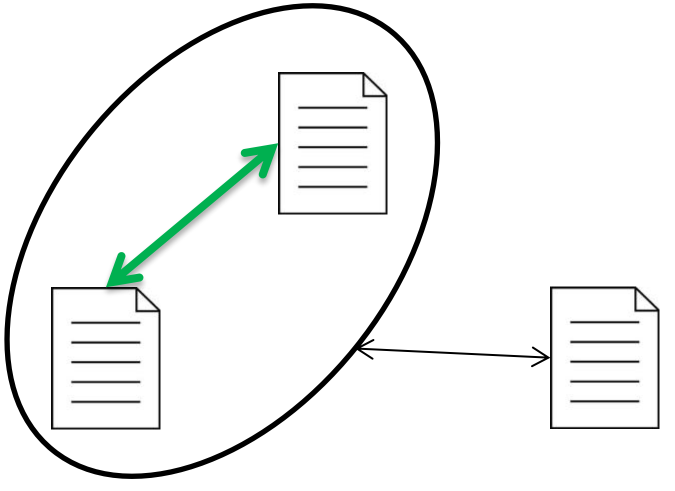
Author Clustering vs. Author Verification

- Any clustering problem can be decomposed into a series of verification problems
  - determine whether any possible pair of documents is by the same author or not.
- Some of these verification problems are strongly correlated
  - this information can be used to enhance the verification accuracy



Author Clustering vs. Author Verification

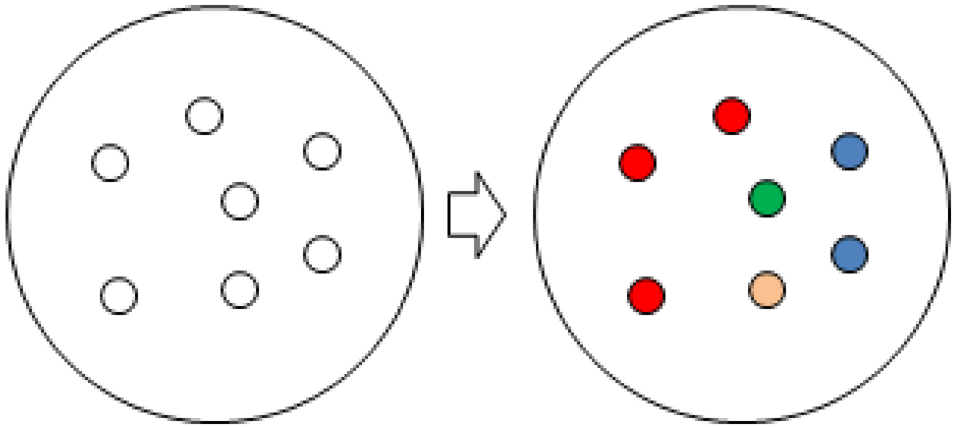
- Any clustering problem can be decomposed into a series of verification problems
  - determine whether any possible pair of documents is by the same author or not.
- Some of these verification problems are strongly correlated
  - this information can be used to enhance the verification accuracy



Task Definition

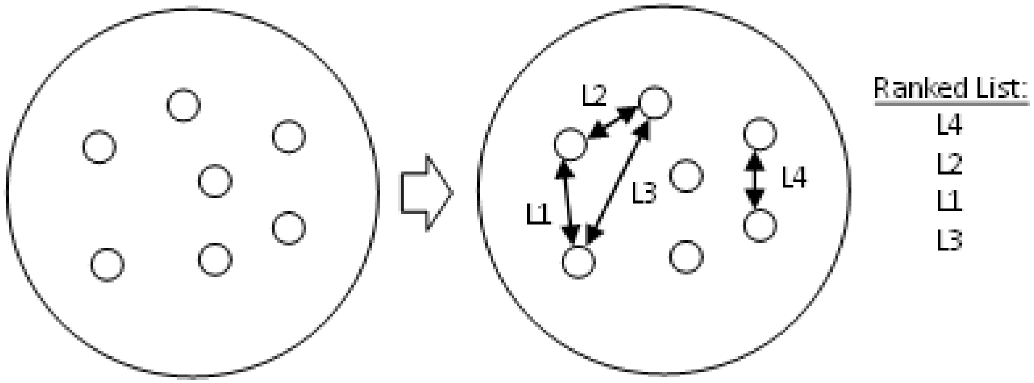
- Given a document collection, group them by authorship and determine all possible authorship links
  - The number of distinct authors is not given
- Assumptions:
  - Each collection comprises up to 100 documents
  - All documents are single-authored
  - All documents are in the same language
  - All documents belong to the same genre
  - The topic of documents may vary
  - The text-length of documents may vary

Complete Author Clustering



- The number of different authors (*k*) found in the collection should be identified
- Each document should be assigned to exactly one of the *k* clusters

Authorship-link Ranking



- Given a document collection, determine authorship links between documents and rank them according to a confidence score
  - Authorship-link: a pair of documents by the same author
  - Confidence score: The higher, the more likely the document pair to be by the same author

Clusteriness Ratio

$$r = k/N$$

- N: the number of documents in the collection
- k the number of distinct authors in this collection
- It indicates:
  - the percentage of single-document clusters
  - the number of available authorship links
- We examine three cases:
  - $r \approx 0.9$: only a few documents belong to multi-document clusters and it is unlikely to find authorship links
  - $r \approx 0.7$: the majority of documents belong to single-document clusters and it is likely to find authorship links
  - $r \approx 0.5$: less than half of the documents belong to single-document clusters and there are plenty of authorship links

PAN-2016 Author Clustering Corpus

- **Dutch articles**: opinion articles from the Flemish daily newspaper *De Standaard* and weekly news magazine *Knack*
- **Dutch reviews**: both positive and negative reviews about both real and fictional products (smartphones, fastfood restaurants, books, artists, and movies) taken from the *CLiPS Stylometry Investigation* corpus
- **English articles**: opinion articles published in *The Guardian* UK daily newspaper
- **English reviews**: book reviews published in *The Guardian* UK daily newspaper
- **Greek articles**: opinion articles published in the online forum www.protagon.gr
- **Greek reviews**: restaurant reviews downloaded from the website www.ask4food.gr

PAN-2016 Author Clustering Corpus

- For each language/genre, three training instances and three test instances:
  - $r \approx 0.9$
  - $r \approx 0.7$
  - $r \approx 0.5$

Training Corpus

id	Language	Genre	r	N	k	Links	maxC	Avg. words
001	English	articles	0.70	50	35	26	5	752.3
002	English	articles	0.50	50	25	75	9	756.2
003	English	articles	0.86	50	43	8	3	744.7
004	English	reviews	0.69	80	55	36	4	977.8
005	English	reviews	0.88	80	70	12	3	1,089.7
006	English	reviews	0.50	80	40	65	5	1,029.4
007	Dutch	articles	0.89	57	51	7	3	1,074.7
008	Dutch	articles	0.49	57	28	76	7	1,321.9
009	Dutch	articles	0.70	57	40	30	4	1,014.8
010	Dutch	reviews	0.54	100	54	77	4	128.2
011	Dutch	reviews	0.67	100	67	46	4	134.9
012	Dutch	reviews	0.91	100	91	10	3	125.3
013	Greek	articles	0.51	55	28	38	4	748.9
014	Greek	articles	0.69	55	38	25	5	741.6
015	Greek	articles	0.87	55	48	8	3	726.8
016	Greek	reviews	0.91	55	50	6	3	523.4
017	Greek	reviews	0.51	55	28	55	8	633.9
018	Greek	reviews	0.73	55	40	19	3	562.9

Test Corpus

id	Language	Genre	r	N	k	Links	maxC	Avg. words
001	English	articles	0.71	70	50	33	5	582.4
002	English	articles	0.50	70	35	113	8	587.3
003	English	articles	0.91	70	64	7	3	579.8
004	English	reviews	0.73	80	58	30	4	1,011.2
005	English	reviews	0.90	80	72	10	3	1,030.4
006	English	reviews	0.53	80	42	68	5	1,003.7
007	Dutch	articles	0.74	57	42	24	4	1,172.1
008	Dutch	articles	0.88	57	50	8	3	1,178.4
009	Dutch	articles	0.53	57	30	65	7	945.2
010	Dutch	reviews	0.88	100	88	16	4	151.7
011	Dutch	reviews	0.51	100	51	76	4	150.3
012	Dutch	reviews	0.71	100	71	37	4	155.9
013	Greek	articles	0.71	70	50	24	4	720.5
014	Greek	articles	0.50	70	35	52	4	750.3
015	Greek	articles	0.89	70	62	9	3	737.6
016	Greek	reviews	0.73	70	51	24	4	434.8
017	Greek	reviews	0.91	70	64	7	3	428.0
018	Greek	reviews	0.53	70	37	44	4	536.9

Evaluation Measures

- Complete author clustering
  - BCubed Precision, Recall, and F-score
  - Extrinsic clustering evaluation
  - They satisfy several formal constraints including cluster homogeneity, cluster completeness, and the *rag bag* criterion
- Authorship-link ranking
  - Mean average precision (official)
  - R-precision
  - P@10

Baselines

- **BASELINE-Random**: based on random guessing
  - The number of authors in a collection is randomly guessed
  - Each document is randomly assigned to one author
  - Authorship links are assigned random scores
  - Average of 50 repetitions for each clustering problem
  - The lower limit of performance
- **BASELINE-Singleton**: all documents belong to different authors
  - All clusters are singleton
  - Very effective when *r* is high
  - It guarantees a BCubed precision of 1
- **BASELINE-Cosine**: determine authorship links based on cosine similarity
  - Text representation: normalized frequencies of all words appearing at least 3 times in the collection
  - It should be affected by topical similarities between documents

Submissions

- We received 8 submissions
  - Bulgaria, India, Iran, New Zealand, Switzerland (2), and UK (2)
- All teams submitted and evaluated their software in TIRA
  - http://www.tira.io/
- 6 notebook submissions

Top-down Approaches

- First attempt to form clusters using a typical clustering algorithm (e.g. *k*-means)
- Then transform clusters into authorship links assigning a score to each link
- Estimating number of authors (*k*) is a crucial decision
  - Sari & Stevenson (2016) use the Silhouette coefficient
  - Mansoorizadeh et al. (2016) use the number of sub-graphs in a document similarity graph

Bottom-up Approaches

- First estimate the pairwise distance of documents (authorship-link scores)
- Then use this information to form clusters
- The number of authors (*k*) is not explicitly estimated
  - Clusters are formed according to certain criteria
  - Kocher (2016) group texts in one cluster if they are connected by a path of authorship links with significantly high score
  - Bagnall (2016) practically forbids clusters with more than two items
- Distance measures are in some cases a modification of author verification approaches
  - Bagnall (2016), Vartapetiance & Gillam (2016)
- Zmiycharov et al. (2016) transform the estimation of authorship link scores to a supervised learning task
  - class imbalance problem

Stylometric Features

- All submissions follow well-known methods
- Homogeneous feature sets:
  - character-level information (Bagnall (2016), Sari & Stevenson (2016))
  - very frequent terms (Kocher (2016), Vartapetiance & Gillam (2016))
- Heterogeneous feature sets:
  - sentence length, type-token ratio, word frequencies, part-of-speech tag frequencies and distributions (Mansoorizadeh et al. (2016), Zmiycharov et al. (2016))
- Sari & Stevenson (2016) report that word embeddings were tested but finally excluded due to low preliminary results

Overall Results

Participant	Complete clustering			Authorship-link ranking			Runtime
	B3 F	B3 rec.	B3 prec.	MAP	RP	P@10	
Bagnall	**0.822**	0.726	0.977	**0.169**	**0.168**	**0.283**	63:03:59
Gobeill	0.706	0.767	0.737	0.115	0.131	0.233	00:00:39
Kocher	**0.822**	0.722	**0.982**	0.054	0.050	0.117	00:01:51
Kuttichira *et al.*	0.588	0.720	0.512	0.001	0.010	0.006	00:00:42
Mansoorizadeh *et al.*	0.401	0.822	0.280	0.009	0.012	0.011	00:00:17
Sari & Stevenson	0.795	0.733	0.893	0.040	0.065	0.217	00:07:48
Vartapetiance & Gillam	0.234	**0.935**	0.195	0.012	0.023	0.044	03:03:13
Zmiycharov *et al.*	0.768	0.716	0.852	0.003	0.016	0.033	01:22:56
BASELINE-Random	0.667	0.714	0.641	0.002	0.009	0.013	–
BASELINE-Singleton	0.821	0.711	**1.000**	–	–	–	–
BASELINE-Cosine	–	–	–	0.060	0.074	0.139	–

Complete Author Clustering Results

- Mean BCubed F-score

Participant	Overall	Articles	Reviews	English	Dutch	Greek	$r \approx 0.9$	$r \approx 0.7$	$r \approx 0.5$
Bagnall	**0.822**	**0.817**	**0.828**	**0.820**	**0.815**	0.832	0.931	0.840	**0.695**
Kocher	**0.822**	**0.817**	0.827	0.818	**0.815**	**0.833**	**0.933**	**0.843**	0.690
BASELINE-Singleton	0.821	**0.819**	0.823	**0.822**	**0.819**	0.822	**0.945**	0.838	0.680
Sari & Stevenson	0.795	0.789	0.801	0.784	0.789	0.813	0.887	0.812	0.687
Zmiycharov *et al.*	0.768	0.761	0.776	0.781	0.759	0.765	0.877	0.777	0.651
Gobeill	0.706	0.800	0.611	0.805	0.606	0.707	0.756	0.722	0.639
BASELINE-Random	0.667	0.666	0.667	0.668	0.665	0.667	0.745	0.678	0.577
Kuttichira *et al.*	0.588	0.626	0.550	0.579	0.584	0.601	0.647	0.599	0.519
Mansoorizadeh *et al.*	0.401	0.367	0.435	0.486	0.256	0.460	0.426	0.373	0.403
Vartapetiance & Gillam	0.234	0.284	0.183	0.057	0.595	0.049	0.230	0.241	0.230

- Number of detected clusters

	N	k	Bagnall	Gobeill	Kocher	Kuttichira *et al.*	Mansoorizadeh *et al.*	Sari & Stevenson	Vartapetiance & Gillam	Zmiycharov *et al.*
problem001	70	50	70	61	68	36	20	60	1	59
problem002	70	35	70	54	68	36	20	60	1	63
problem003	70	64	70	56	68	36	20	60	1	60
problem004	80	58	80	77	78	36	25	70	1	73
problem005	80	72	79	78	78	36	31	70	1	74
problem006	80	42	78	78	77	36	29	70	1	71
problem007	57	42	54	50	55	36	1	48	42	47
problem008	57	50	55	48	55	36	11	48	39	49
problem009	57	30	56	49	55	36	2	48	46	49
problem010	100	88	99	28	97	36	20	90	28	84
problem011	100	51	96	23	98	36	20	90	25	86
problem012	100	71	98	29	98	36	20	90	33	80
problem013	70	50	69	61	68	36	20	60	1	55
problem014	70	35	70	63	68	36	20	60	1	59
problem015	70	62	70	66	66	36	20	60	1	58
problem016	70	51	56	29	67	36	20	60	1	58
problem017	70	64	59	23	68	36	20	60	1	58
problem018	70	37	58	31	67	36	20	60	1	53

Authorship-link Ranking Results

- Mean Average Precision

Participant	Overall	Articles	Reviews	English	Dutch	Greek	$r\approx0.9$	$r\approx0.7$	$r\approx0.5$
Bagnall	**0.169**	**0.174**	**0.163**	**0.126**	**0.109**	**0.272**	**0.064**	**0.186**	**0.257**
Gobeill	0.115	0.119	0.110	0.097	0.079	0.168	0.040	0.105	0.198
BASELINE-Cosine	0.060	0.063	0.057	0.053	0.053	0.074	0.019	0.054	0.107
Kocher	0.054	0.047	0.061	0.032	0.044	0.085	0.042	0.058	0.063
Sari & Stevenson	0.040	0.033	0.047	0.009	0.042	0.069	0.017	0.041	0.062
Vartapetiance & Gillam	0.012	0.010	0.014	0.014	0.006	0.016	0.010	0.008	0.017
Mansoorizadeh *et al.*	0.009	0.013	0.004	0.006	0.010	0.010	0.002	0.009	0.014
Zmiycharov *et al.*	0.003	0.002	0.004	0.001	0.000	0.009	0.002	0.003	0.004
BASELINE-Random	0.002	0.002	0.001	0.001	0.002	0.002	0.001	0.001	0.002
Kuttichira *et al.*	0.001	0.002	0.001	0.001	0.002	0.001	0.001	0.002	0.001

- Number of detected authorship links

	true links	max links	Bagnall	Gobeill	Kocher	Kuttichira *et al.*	Mansoorizadeh *et al.*	Sari & Stevenson	Vartapetiance & Gillam	Zmiycharov *et al.*
problem001	33	2415	2415	2415	2415	68	170	14	526	19
problem002	113	2415	2415	2415	2415	57	189	11	529	18
problem003	7	2415	2415	2415	2415	67	262	13	611	16
problem004	30	3160	3160	3160	3160	120	605	23	2705	11
problem005	10	3160	3160	3160	3160	126	614	18	2750	9
problem006	68	3160	3160	3160	3160	88	605	21	2691	10
problem007	24	1596	1596	1596	1596	52	1596	11	36	18
problem008	8	1596	1596	1596	1596	42	475	11	40	23
problem009	65	1596	1596	1596	1596	51	1486	30	21	24
problem010	16	4950	4950	4950	4950	214	323	11	94	79
problem011	76	4950	4950	4950	4950	261	464	14	107	98
problem012	37	4950	4950	4950	4950	229	297	13	91	97
problem013	24	2415	2415	2415	2415	62	288	12	616	94
problem014	52	2415	2415	2415	2415	114	444	13	642	104
problem015	9	2415	2415	2415	2415	70	335	13	833	95
problem016	24	2415	2415	2415	2415	108	335	14	954	36
problem017	7	2415	2415	2415	2415	96	932	23	865	30
problem018	44	2415	2415	2415	2415	87	859	23	1134	51

Conclusions

- First shared task in unsupervised authorship analysis
  - Author clustering is a challenging task
- Clusteriness ratio *r* represents both the quantity of authorship links and the number of single-item clusters
- Few submissions were able to surpass BASELINE-Singleton
- Few submissions were able to surpass BASELINE-Cosine
- Best results were achieved by a modification of an author verification method
  - Author clustering and author verification are strongly related tasks
- Bottom-up approaches seem to be more effective
- Homogeneous feature sets seem to be more suitable

Future Work

- Focus on short texts
  - Paragraph-length
  - Tweets
- Drop the assumption that all documents belong to the same genre
- Consider documents from distant thematic areas