Overview of the Author Identification Task at PAN 2014

Efstathios Stamatatos

University of the Aegean

Walter Daelemans

University of Antwerp

Ben Verhoeven

University of Antwerp

Martin Potthast

Bauhaus-Universität Weimar

Benno Stein

Bauhaus-Universität Weimar

Patrick Juola

Duquesne University

Miguel Angel Sánchez Pérez

National Polytechnic Institute, Mexico

Alberto Barrón-Cedeño

Universitat Politècnica de Catalunya

Outline

- Introduction
- Evaluation setup
- Evaluation results
- Survey of submissions
- Conclusions

Authorship Analysis

- Author identification: Given a set of candidate authors for whom some texts of undisputed authorship exist, attribute texts of unknown authorship to one of the candidates
- Author profiling: The extraction of demographic information such as gender, age, etc. about the authors
- Author clustering: The segmentation of texts into stylistically homogeneous parts

Author Identification Tasks

- Closed-set: there are several candidate authors, each represented by a set of training data, and one of these candidate authors is assumed to be the author of unknown document(s)
- Open-set: the set of potential authors is an open class, and “none of the above” is a potential answer
- Authorship verification: the set of candidate authors is a singleton and either he wrote the unknown document(s) or “someone else” did

Evaluation Setup

- *Given a set of documents (no more than 5, possibly only one) by the same author, is an additional (out-of-set) document also by that author?*
- All the documents within a verification problem are matched in language, genre, theme, and date of writing
- Participants were asked to produce a real score in [0,1] inclusive, where
  - 1.0 corresponds to "certainly yes"
  - 0.0 corresponds to "certainly no"
  - 0.5 corresponds to "I don't know"
- Software submissions were required

Author Identification Task at PAN-2013 vs. PAN-2014

- Similarities with PAN-2013
  - Same task definition
  - Software submissions required
  - Corpora in several languages
- Differences with PAN-2013
  - Real scores are obligatory
  - Real scores should be calibrated
  - Larger corpora are provided
  - Richer set of languages and genres
  - More appropriate evaluation measures

PAN-2014 Corpus

	Language	Genre	#Problems	#Docs	Avg. of known docs per problem	Avg. words per document
Training	Dutch	Essays	96	268	1.8	412.4
	Dutch	Reviews	100	202	1.0	112.3
	English	Essays	200	729	2.6	848.0
	English	Novels	100	200	1.0	3,137.8
	Greek	Articles	100	385	2.9	1,404.0
	Spanish	Articles	100	600	5.0	1,135.6
	Total		**696**	**2,384**	**2.4**	**1,091.0**
Evaluation	Dutch	Essays	96	287	2.0	398.1
	Dutch	Reviews	100	202	1.0	116.3
	English	Essays	200	718	2.6	833.2
	English	Novels	200	400	1.0	6,104.0
	Greek	Articles	100	368	2.7	1,536.6
	Spanish	Articles	100	600	5.0	1,121.4
	Total		**796**	**2,575**	**2.2**	**1,714.9**
TOTAL			**1,492**	**4,959**	**2.3**	**1,415.0**

PAN-2014 Corpus

- The Dutch corpus is a transformed version of the CLiPS Stylometry Investigation (CSI) corpus
  - All documents by language students at the University of Antwerp between 2012 and 2014
- The English essays corpus was derived from the Uppsala Student English (USE) corpus
  - All documents by English-as-second-language students
- The English novels corpus focuses on a very small subgenre of speculative and horror fiction known as the “Cthulhu Mythos” (“Lovecraftian horror”)
  - Documents were gathered from a variety of on-line sources including Project Gutenberg and FanFiction

PAN-2014 Corpus

- The Greek corpus comprises newspaper opinion articles published in the Greek weekly newspaper *TO BHMA* from 1996 to 2012
  - In contrast to PAN-2013 only thematic similarities were used to form verification problems
- The Spanish corpus includes newspaper opinion articles of the Spanish newspaper *El Pais*
  - Verification problems were formed taking into account thematic similarities between articles

- All corpora are balanced (positive/negative problems)

Performance Measures

- AUC of ROC curves
- c@1

$$\mathrm{c@1} = \frac{1}{\mathrm{n}}\left(\mathrm{n_c} + \frac{\mathrm{n_c}}{\mathrm{n}}\mathrm{n_u}\right)$$

  - able to take unanswered problems into account
  - explicitly extends accuracy based on the number of problems left unanswered
  - originally proposed for question answering tasks

- The final rank of participants is based on the product of AUC and c@1
- Efficiency is measured by elapsed runtime

Baseline

- Instead of using random guessing, we adopted a more challenging baseline that can reflect and adapt to the difficulty of a specific corpus
- [Jankowska et al., 2013]
  - It is language-independent
  - It can provide both binary answers and real scores
  - The real scores are already calibrated to probability-like scores for a positive answer
  - It was the winner of PAN-2013 in terms of overall AUC scores
- It has not been specifically trained on the corpora of PAN-2014
- Not able to leave problems unanswered

Meta-classifier

- A meta-model that combines all answers given by the participants for each problem
  - the average of the probability scores provided by the participants for each problem
  - Not tuned to leave more problems unanswered
- Similar idea with PAN-2013
  - Heterogeneous models seem to be very effective

Submissions

- We received 13 submissions
  - from research teams in Australia, Canada (2), France, Germany (2), India, Iran, Ireland, Mexico (2), United Arab Emirates, and United Kingdom
- The participants submitted and evaluated their software within the TIRA framework
- A separate run for each corpus corresponding to each language and genre was performed

Overall Results (micro-averaging)

Rank		FinalScore	AUC	c@1	Runtime	Unanswered Problems
	META-CLASSIFIER	0.566	0.798	0.710		0
1	Khonji & Iraqi	0.490	0.718	0.683	20:59:40	2
2	Frery et al.	0.484	0.707	0.684	00:06:42	28
3	Castillo et al.	0.461	0.682	0.676	03:59:04	78
4	Moreau et al.	0.451	0.703	0.641	01:07:34	50
5	Mayor et al.	0.450	0.690	0.651	05:26:17	29
6	Zamani et al.	0.426	0.682	0.624	02:37:25	0
7	Satyam et al.	0.400	0.631	0.634	02:52:37	7
8	Modaresi & Gross	0.375	0.610	0.614	00:00:38	0
9	Jankowska et al.	0.367	0.609	0.602	07:38:18	7
10	Halvani & Steinebach	0.335	0.595	0.564	00:00:54	3
	BASELINE	0.325	0.587	0.554	00:21:10	0
11	Vartapetiance & Gillam	0.308	0.555	0.555	01:07:39	0
12	Layton	0.306	0.548	0.559	27:00:01	0
13	Harvey	0.304	0.558	0.544	01:06:19	100

Results on Dutch Essays

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
META-CLASSIFIER	0.867	0.957	0.906		0
Mayor et al.	0.823	0.932	0.883	00:15:05	2
Frery et al.	0.821	0.906	0.906	00:00:30	0
Khonji & Iraqi	0.770	0.913	0.844	00:58:21	0
Moreau et al.	0.755	0.907	0.832	00:02:09	34
Castillo et al.	0.741	0.861	0.861	00:01:57	2
Jankowska et al.	0.732	0.869	0.842	00:23:26	1
BASELINE	0.685	0.865	0.792	00:00:52	0
Zamani et al.	0.525	0.741	0.708	00:00:27	0
Vartapetiance & Gillam	0.517	0.719	0.719	00:06:37	0
Satyam et al.	0.489	0.651	0.750	00:01:21	0
Halvani & Steinebach	0.399	0.647	0.617	00:00:06	2
Harvey	0.396	0.644	0.615	00:02:19	0
Modaresi & Gross	0.378	0.595	0.635	00:00:05	0
Layton	0.307	0.546	0.563	00:55:07	0

Results on Dutch Reviews

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
Satyam et al.	0.525	0.757	0.694	00:00:16	2
Khonji & Iraqi	0.479	0.736	0.650	00:12:24	0
META-CLASSIFIER	0.428	0.737	0.580		0
Moreau et al.	0.375	0.635	0.590	00:01:25	0
Zamani et al.	0.362	0.613	0.590	00:00:11	0
Jankowska et al.	0.357	0.638	0.560	00:06:24	0
Frery et al.	0.347	0.601	0.578	00:00:09	5
BASELINE	0.322	0.607	0.530	00:00:12	0
Halvani & Steinebach	0.316	0.575	0.550	00:00:03	0
Mayor et al.	0.299	0.569	0.525	00:07:01	1
Layton	0.261	0.503	0.520	00:56:17	0
Vartapetiance & Gillam	0.260	0.510	0.510	00:05:43	0
Castillo et al.	0.247	0.669	0.370	00:01:01	76
Modaresi & Gross	0.247	0.494	0.500	00:00:07	0
Harvey	0.170	0.354	0.480	00:01:45	0

Results on English Essays

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
META-CLASSIFIER	0.531	0.781	0.680		0
Frery et al.	0.513	0.723	0.710	00:00:54	15
Satyam et al.	0.459	0.699	0.657	00:16:23	2
Moreau et al.	0.372	0.620	0.600	00:28:15	0
Layton	0.363	0.595	0.610	07:42:45	0
Modaresi & Gross	0.350	0.603	0.580	00:00:07	0
Khonji & Iraqi	0.349	0.599	0.583	09:10:01	1
Halvani & Steinebach	0.338	0.629	0.538	00:00:07	1
Zamani et al.	0.322	0.585	0.550	00:02:03	0
Mayor et al.	0.318	0.572	0.557	01:01:07	10
Castillo et al.	0.318	0.549	0.580	01:31:53	0
Harvey	0.312	0.579	0.540	00:10:22	0
BASELINE	0.288	0.543	0.530	00:03:29	0
Jankowska et al.	0.284	0.518	0.548	01:16:35	5
Vartapetiance & Gillam	0.270	0.520	0.520	00:16:44	0

Results on English Novels

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
Modaresi & Gross	0.508	0.711	0.715	00:00:07	0
Zamani et al.	0.476	0.733	0.650	02:02:02	0
META-CLASSIFIER	0.472	0.732	0.645		0
Khonji & Iraqi	0.458	0.750	0.610	02:06:16	0
Mayor et al.	0.407	0.664	0.614	01:59:47	8
Castillo et al.	0.386	0.628	0.615	02:14:11	0
Satyam et al.	0.380	0.657	0.579	02:14:28	3
Frery et al.	0.360	0.612	0.588	00:03:11	1
Moreau et al.	0.313	0.597	0.525	00:11:04	12
Halvani & Steinebach	0.293	0.569	0.515	00:00:07	0
Harvey	0.283	0.540	0.525	00:46:30	0
Layton	0.260	0.510	0.510	07:27:58	0
Vartapetiance & Gillam	0.245	0.495	0.495	00:13:03	0
Jankowska et al.	0.225	0.491	0.457	02:36:12	1
BASELINE	0.202	0.453	0.445	00:08:31	0

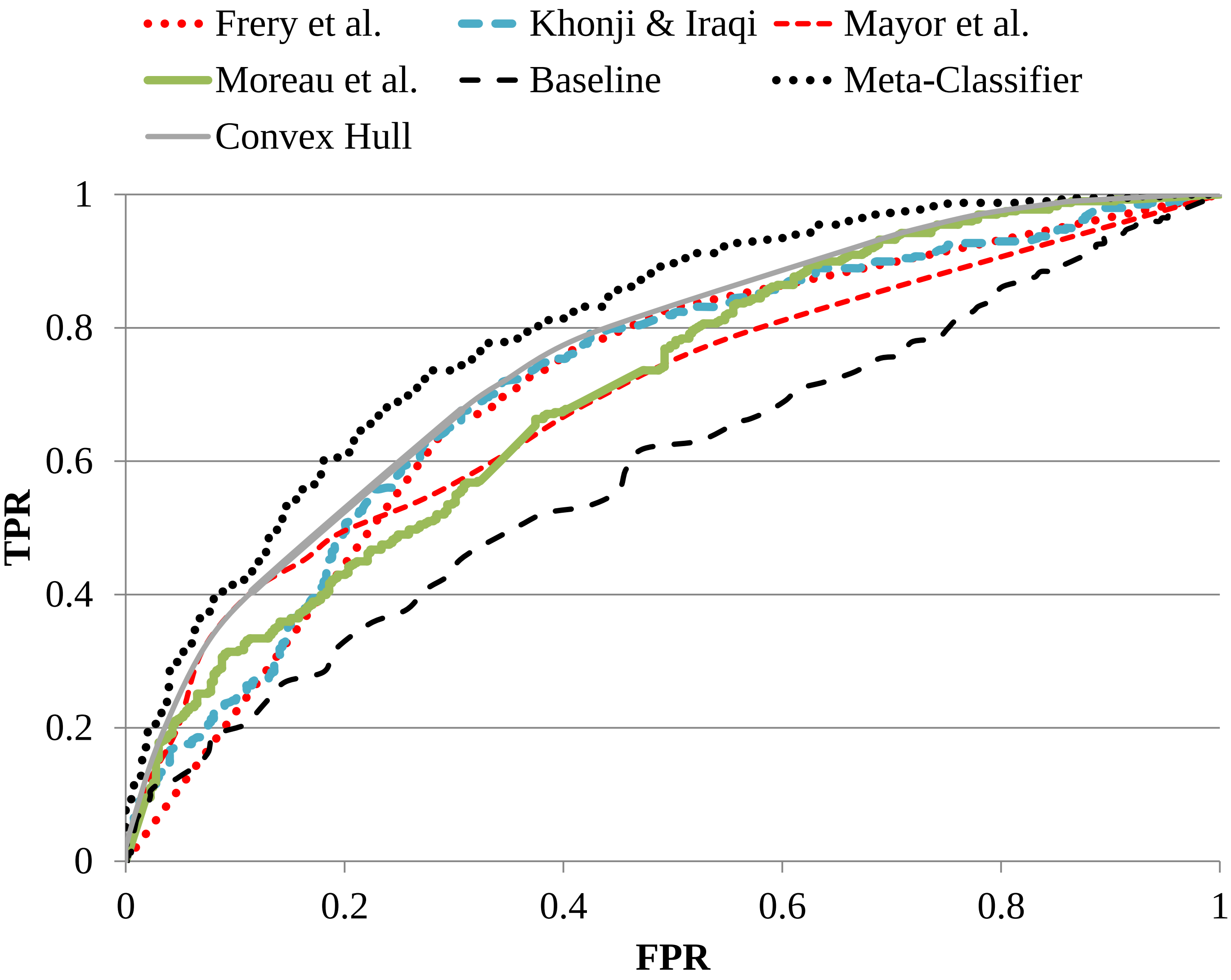
Results on Greek Articles

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
Khonji & Iraqi	0.720	0.889	0.810	03:41:48	0
META-CLASSIFIER	0.635	0.836	0.760		0
Mayor et al.	0.621	0.826	0.752	00:51:03	3
Moreau et al.	0.565	0.800	0.707	00:05:54	4
Castillo et al.	0.501	0.686	0.730	00:03:14	0
Jankowska et al.	0.497	0.731	0.680	01:36:00	0
Zamani et al.	0.470	0.712	0.660	00:15:12	0
BASELINE	0.452	0.706	0.640	00:03:38	0
Frery et al.	0.436	0.679	0.642	00:00:58	7
Layton	0.403	0.661	0.610	04:40:29	0
Halvani & Steinebach	0.367	0.611	0.600	00:00:04	0
Satyam et al.	0.356	0.593	0.600	00:12:01	0
Modaresi & Gross	0.294	0.544	0.540	00:00:05	0
Vartapetiance & Gillam	0.281	0.530	0.530	00:10:17	0
Harvey	0.000	0.500	0.000		100

Results on Spanish Articles

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
META-CLASSIFIER	0.709	0.898	0.790		0
Khonji & Iraqi	0.698	0.898	0.778	04:50:49	1
Moreau et al.	0.634	0.845	0.750	00:18:47	0
Jankowska et al.	0.586	0.803	0.730	01:39:41	0
Frery et al.	0.581	0.774	0.750	00:01:01	0
Castillo et al.	0.558	0.734	0.760	00:06:48	0
Mayor et al.	0.539	0.755	0.714	01:12:14	5
Harvey	0.514	0.790	0.650	00:05:23	0
Zamani et al.	0.468	0.731	0.640	00:17:30	0
Vartapetiance & Gillam	0.436	0.660	0.660	00:15:15	0
Halvani & Steinebach	0.423	0.661	0.640	00:00:27	0
Modaresi & Gross	0.416	0.640	0.650	00:00:08	0
BASELINE	0.378	0.713	0.530	00:04:27	0
Layton	0.299	0.553	0.540	05:17:25	0
Satyam et al.	0.248	0.443	0.560	00:08:09	0

ROC Curves (overall)



Statistical Significance Test

- We computed statistical significance of performance differences between systems using approximate randomization testing (ART)
- Paired t-tests make assumptions that do not hold for precision scores and F-scores
- ART does not make these assumptions and can handle complicated distributions
- The null hypothesis is that there is no difference in the output of two systems

Results of Statistical Significance Tests

	Khonji & Iraqi	Frery et al.	Castillo et al.	Moreau et al.	Mayor et al.	Zamani et al.	Satyam et al.	Modaresi & Gross	Jankowska et al.	Halvani & Steinebach	BASELINE	Vartapetiance & Gillam	Layton	Harvey
META-CLASSIFIER	=	*	***	***	***	***	***	***	***	***	***	***	***	***
Khonji & Iraqi		=	**	***	**	**	**	**	**	***	***	***	***	***
Frery et al.			=	*	=	=	=	=	*	***	***	***	**	***
Castillo et al.				=	=	=	=	=	=	*	**	**	*	***
Moreau et al.					=	=	=	=	=	=	*	*	=	***
Mayor et al.						=	=	=	=	**	**	**	**	***
Zamani et al.							=	=	=	**	**	**	**	***
Satyam et al.								=	=	**	**	***	**	***
Modaresi & Gross									=	*	**	*	*	***
Jankowska et al.										=	**	=	=	***
Halvani & Steinebach											=	=	=	***
BASELINE												=	=	*
Vartapetiance & Gillam													=	**
Layton														**

Survey of Submissions: Intrinsic vs. extrinsic verification

- Intrinsic methods use only the known texts and the unknown text of a problem
  - The majority of submitted approaches
- External methods make use of additional texts by other authors
  - Transform author verification from a one-class to a binary classification task
  - The winner of PAN-2014 is a modification of the *Impostors* method [Koppel & Winter, 2014], similarly to PAN-2013
  - Other external approaches are described by [Mayor et al.] and [Zamani et al.]

Survey of Submissions: Type of learning

- In lazy approaches the training phase is nearly omitted and all necessary processing is performed at the time they have to decide about a new verification problem
  - Most of the submitted approaches follow this idea
  - All PAN-2013 submissions as well
- Eager methods attempt to build a general model based on the training corpus
  - decision trees [Frery et al.], genetic algorithms [Moreau et al.], fuzzy C-means clustering [Modaresi & Gross]
- PAN-2014 corpus size permits this type of learning
- Eager methods are generally more efficient

Survey of Submissions: Text representation

- The majority of the participant methods focused on low-level measures
  - character measures (i.e., punctuation mark counts, prefix/suffix counts, character n-grams, etc.)
  - lexical measures (i.e., vocabulary richness measures, sentence/word length counts, stopword frequency, n-grams of words/stopwords, word skip-grams, etc.)
- Only a few attempts to incorporate syntactic features
  - POS tag counts [Khonji & Iraqi], [Moreau et al.] [Zamani et al], [Harvey]

Second-time Participants

- In total 13 participant approaches
  - 7 were also participated in PAN-2013
  - Some attempted to improve the method proposed in 2013 and others presented new models

- Remarkably those teams that slightly modified their existing approach did not achieve a high performance
  - [Halvani & Steinebach], [Jankowska et al] [Layton] [Vartapetiance & Gillam]
- The teams that radically changed their approach, including the ability to leave some problems unanswered, achieved very good results
  - [Castillo et al], [Mayor et al.], [Moreau et al.]

Conclusions

- PAN-2014 Corpora are substantially enlarged
  - Including several languages and genres
- Participants enabled to study how to adapt and fine-tune their approaches for a given language and genre
- Use of different performance measures that put emphasis on both
  - the appropriate ranking of the provided answers in terms of confidence (AUC)
  - the ability of the submitted systems to leave some problems unanswered when there is great uncertainty (c@1)

Conclusions

- Similar to PAN-2013, the overall winner was a modification of the *Impostors* method
  - great potential of extrinsic verification models
- The significantly larger training corpus allowed participants to explore, for the first time, the use of eager learning methods in author verification
  - both effective and efficient

Conclusions

- A challenging baseline method was used
  - A PAN-2013 participant
  - Better baseline methods can be used in future competitions
- The meta-classifier combining all submitted systems in a heterogeneous ensemble was better than each individual submitted method
  - its ROC curve clearly outperformed the convex hull of all submitted approaches.
  - great potential of heterogeneous models in author verification
- Statistical significance tests reveal that there is no significant difference between systems ranked in neighboring positions
  - However, the winner approach is significantly better than the rest of the submissions (excluding the second winner)

Future Work

- The focus of PAN-2013 and PAN-2014 on the author verification task has produced a significant progress in this field
  - Development of new corpora
  - Development of new methods
  - Defining an appropriate evaluation framework
- Author verification is far from being a solved task
  - There are many variations that can be explored in future evaluation labs
  - Cross-topic verification (the known and the questioned documents do not match in terms of topic)
  - Cross-genre verification (the known and the questioned documents do not match in terms of genre)
  - Any comments/suggestions are welcome