Overview of the International Authorship Identification Competition at PAN-2011

Shlomo Argamon[1] and Patrick Juola[2]

[1] Linguistic Cognition Lab, Department of Computer Science
Illinois Institute of Technology, Chicago, IL 60616
argamon@iit.edu
[2] Department of Mathematics and Computer Science
Duquesne University, Pittsburgh, PA 15282
juola@mathcs.duq.edu

Abstract. This paper gives an overview of the evaluation methodology applied to authorship identification solutions as part of PAN 2011. The two variations of authorship identification that were explored were *authorship attribution*, determining which of a known set of authors wrote a text, and *authorship verification*, determining if a specific authors did or did not write a text. We summarize the methods used by the various participants, which were quite varied, and present the overall results of the evaluation.

1 Introduction

There has been much interest in recent years in research on automatic methods for determining the authorship of anonymous documents based on internal evidence [7, 16, 9]. Indeed, accurate automatic authorship attribution of anonymous documents is of increasing importance for many applications, including homeland security, criminal and civil law, computer forensics, and humanities scholarship. However, despite the growing need for effective and reliable methods, research has been hampered by the lack of any canonical testbed for authorship attribution. Combined with the interdisciplinary nature of the field, this has often led to redundant and unsound research. The purpose of this authorship competition, held as part of the 2011 PAN Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse, is to start redressing this problem, by advancing a standardized evaluation framework for authorship attribution and related problems.

A total of 13 different research groups submitted results for 7 different tasks within this evaluation framework, eight of which submitted papers describing their systems. In this paper we describe the evaluation framework, and report on evaluation of the different authorship analysis methods.

1.1 The problem

In the basic form of the *authorship attribution problem*, we are given examples of the writing of a number of candidate authors and are asked to determine

which of them authored a given anonymous text. In this straightforward form, the authorship attribution problem fits the standard modern paradigm of a text categorization problem [15]. The components of text categorization systems are by now fairly well-understood: documents are represented as numerical vectors that capture statistics of potentially relevant features of the text and machine learning methods are used to find classifiers that separate documents that belong to different classes.

However, real-life authorship identification problems are rarely as elegant as straightforward "research-type" text categorization problems, in which we have a small closed set of candidate authors and essentially unlimited training text for each. One important issue that arises in the real world is the existence of an open candidate set, that is, the actual author might be an author we don't know about at all. In this case, the problem is to assign the document either to one of the authors we know of, or to "Someone Else".

The most reduced version of this open-candidate case is that where there is no candidate set at all, but just a single suspect. In this case, the challenge is to determine if the suspect is or is not the author. This is called the *authorship verification problem*. As a categorization problem, verification is significantly more difficult than basic attribution and less work has been done on it, but see, e.g., [18, 8, 11, 5]. If, say, we just need to know if a text was written by Shakespeare or by Marlowe, we could just compare the candidate against their respective known texts. If, however, we needed to know if the text was written by Shakespeare or anyone else, it would be difficult to assemble a sufficiently representative sample of non-Shakespeare texts to compare against, and something more sophisticated would be required.

2 Evaluation

2.1 Corpus

A corpus was developed, based on the Enron email corpus[3], to account for several different common attribution and verification scenarios. The corpus contains five separate training collections, and seven test collections, as follows. Two training sets are provided for authorship attribution, a "Large" set containing 9337 documents by 72 different authors and a "Small" set containing 3001 documents by 26 different authors (the author sets are disjoint). For each attribution problem, two test sets are provided, one containing texts only written by the authors in the training set, and one also containing texts written by around 20 other authors each.

The other three training sets are for verification, and so contain only emails from a single author (different from those in other training sets). The verification training sets contain 42, 55, and 47 documents, respectively. Each has an associated test set comprising a mixture of documents written by the training

[3] http://www.cs.cmu.edu/ enron/

author and written by others (some of these are from the Enron corpus, and some are not).

As the tasks are intended to reflect a natural task environment, there are some texts, both in training and in testing sets, that are not in English, or that are automatically generated.

Personal names and email addresses in the corpus have been automatically redacted, and replaced (on a token-by-token basis) by ¡NAME/¿ and ¡EMAIL/¿ tags, respectively. This redaction is admittedly imperfect, but random spot-checking was applied to reduce the likelihood of missing occurrences. Other than this redaction, each text is typographically identical to the original electronic text, so systems could, in principle, rely on line length, punctuation, and the like.

Finally, authorship was determined based on `From:` email headers; this necessitated determining, in some cases, that multiple email addresses corresponded to the same individual. Manual spot-checking was applied here as well to ensure quality, though some errors were let through and discovered during the evaluation.

2.2 Metrics

For evaluating authorship identification, we used the standard information retrieval metrics of precision, recall, and F1. Precision, for a particular author A, is defined as the fraction of attributions that a system makes to A that are correct:

$$P_A = \frac{\text{correct}(A)}{\text{attributions}(A)}$$

Recall, for a particular author A, is defined as the fraction of test documents written by A that are (correctly) attributed to A:

$$R_A = \frac{\text{correct}(A)}{\text{documents-by}(A)}$$

F_1 is defined as the harmonic mean of recall and precision:

$$F_1 = \frac{2P_A R_A}{P_A + R_A}$$

For the authorship attribution tasks, we need to aggregate these measures over all the different test authors. We applied two methods with different properties, macro-averaging and micro-averaging. For a given metric M, set of n authors $\{A_i\}$, with a total of k test documents, these are defined as:

$$\text{macro-avg}_M(\{A_i\}) = \tfrac{1}{n} \textstyle\sum_i M_{A_i}$$

$$\text{micro-avg}_M(\{A_i\}) = \tfrac{1}{k} \textstyle\sum_i k_i M_{A_i}$$

where k_i is the number of test documents written by author A_i. Micro-averaging will give more credit to accuracy on authors with more test documents, while

macro-averaging gives the same credit to all authors, even if they wrote just one test document.

For authorship verification, the author set contains just one author, so averaging is not necessary.

Finally, to achieve an overall ranking for each task, we ranked system performances for each of the measures–six for attribution (macro- and micro-averaged P, R, and F_1) and three for verification, and summed the ranks for each entry in each task. The lower the rank sum, the better (overall) the performance.

3 Survey of Submissions

All documented submissions, within their diversity, followed classic methodology for authorship identification, and (a) identified a set of features that were calculated from the texts, whose values then (b) served as input to some classification algorithm. In this section, we summarize the submissions in terms of what features they used and what algorithms they used. We note that, as in any such summary of varied systems, we inevitably must oversimplify, so please see the full papers describing each system for more information.

3.1 Features

There were a number of different kinds of features used by participants, some traditional, some quite novel.

The first type of feature are those derived from the word usage in the texts, which we term *lexical* features. Simplest, are the frequencies of the various words and word n-grams that appear in the text. Also relevant, based on previous studies, are the relative frequencies of function words (or stopwords) and of specific classes of words: pronouns, modal verbs, discourse linking words/phrases (such as "however", "on the other hand"), slang terms, contractions, and emoticons (or smileys). Also considered were frequencies of US vs. UK variants (a dialect indicator), various types of spelling errors, different types of named entities (people, organizations, dates, etc.), and semantic features of words (polysemy, specificity of meaning, etc.).

The second type of feature are those at the character level, and include character n-grams (usually for $n = 3$), frequent suffixes, and punctuation usage.

The third type of feature considered relate to the format of the text, including various length-related features (lengths of lines, words, sentences), overall formatting of the text (e.g., fraction of empty lines), orthographic features (e.g., capitalization, frequency of non-alphanumeric characters), and a novel feature, Intro/Outro that looked for common beginnings and endings of texts, and noted their presence/absence as cues to authorship.

The fourth type of feature were syntax related features, both part-of-speech n-grams and phrase types (or dependency link types).

Additionally, two of the submissions used forms of complexity measures over sentences and words, by measuring such things as perplexity and morphological complexity.

Lexical features
Words: Solorio et al., Vilario et al., Mikros & Perifanos, Luyckx, Kern et al., Tanguy et al.
Word n-grams: Mikros & Perifanos, Luyckx, Kern et al.
Function words: Kern et al., Tanguy et al.
Pronouns: Kern et al.
Discourse words: Luyckx
Modal verbs: Luyckx
Slang: Kern et al.
Contractions/abbreviations: Solorio et al., Tanguy et al.
Emoticons: Kern et al., Tanguy et al.
Spelling error types: Tanguy et al.
US/UK variants: Tanguy et al.
Semantics (polysemy, specificity): Tanguy et al.
Named entity types: Tanguy et al.
Character Features
Character n-grams: Kouris & Stamatatos, Mikros & Perifanos, Luyckx, Escalante et al., Tanguy et al.
Suffixes: Tanguy et al.
Punctuation: Solorio et al., Kern et al., Tanguy et al.
Formatting Features
Length (of text, sentence, words): Solorio et al., Kern et al., Tanguy et al.
Text formatting: Kern et al.
Orthography (capitalization, etc.): Solorio et al., Kern et al., Tanguy et al.
Intro/Outro features: Kern et al.
Syntactic Features
Parts of speech: Solorio et al.
Syntax (dependencies, phrase types): Solorio et al., Kern et al.
Others
Complexity measures: Solorio et al., Tanguy et al.
Cluster centroid distances: Solorio et al.

Fig. 1. Feature types and submissions using them.

The last, perhaps most novel, kind of feature used was one used by Solorio et al., based on clustering the training data and measuring the distance of various texts from the cluster centroids, using those distances as features for learning.

The types of features and the submissions using each are listed in Figure 1.

3.2 Algorithms

A wide variety of algorithmic approaches were taken by the participants. Several used different forms of linear classifiers. Support vector machines [4] were used by Solorio et al., and a variant for multiclass problems, $\text{SVM}^{\text{multiclass}}$ [17] was used by Luyckx. Vilariño compared three approaches: the linear Rocchio [14] and Naive Bayes [10] methods, and 100-nearest neighbor [3]. Mikros and Perifanos used the RLR logistic regression algorithm [6].

Other machine learning approaches were also applied. Tanguy et al. applied maximum entropy learning [12] for attribution, and decision trees [13] and rule learning [2] for verification. Kouris and Stamatatos used a co-training approach, combining a kind of nearest-neighbor classifier with a support vector machine approach, to label unlabeled data to improve training. Escalante used a unique form of ensemble learning, EPSMS [5]. Finally, Kern et al. applied a complex multi-level learning scheme using base classifiers which were either bagged decision forests [1] or support vector machines, depending on the feature types, and a probabilistic metaclassifier to integrate base classifications.

	Macro-averaged			**Micro-averaged**			**Rank**
Run	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**	**Sum**
kourtis-2011-06-08-1000	0.549	0.532	0.52	0.658	0.658	0.658	12
kern-2011-06-08-1500	0.615	0.442	0.465	0.642	0.642	0.642	17
tanguy-2011-06-07-1600	0.62	0.444	0.459	0.594	0.594	0.594	17
tanguy-2011-06-07-1700	0.62	0.444	0.459	0.594	0.594	0.594	17
snider-2011-06-08-1548	0.714	0.321	0.384	0.7	0.482	0.571	30
mikros-2011-06-08-2245	0.391	0.356	0.353	0.519	0.519	0.519	41
luyckx-2011-06-10-1640	0.391	0.344	0.342	0.522	0.522	0.522	41
escalante-2011-06-07-0934	0.608	0.294	0.303	0.508	0.508	0.508	48
luyckx-2011-06-10-1635	0.348	0.345	0.34	0.5	0.5	0.5	53
vilarino-2011-05-31-1456	0.364	0.337	0.364	0.428	0.428	0.428	55
vilarino-2011-05-31-1455	0.534	0.095	0.103	0.238	0.238	0.238	69
ryan-2011-06-08-2331	0.186	0.19	0.172	0.255	0.255	0.255	70
eriksson-2011-06-13-0920	0.508	0.094	0.1	0.221	0.221	0.221	75
vilarino-2011-05-31-1454	0.232	0.139	0.147	0.219	0.219	0.219	79
solorio-2011-06-08-1217	0.171	0.084	0.066	0.148	0.148	0.148	91
noecker-2011-06-08-2356	0.231	0.041	0.057	0.035	0.035	0.035	94

Fig. 2. Results for the Large test set without extraneous documents.

4 Evaluation Results

4.1 Attribution

Authorship attribution results, for the four attribution tasks, are given in Tables 2 through 5. As the tables show, the authorship attribution approach of Tanguy et al. was very highly ranked across all the attribution tasks. It was beaten only once significantly by the approach of Kourtis and Stamatatos on the Large task. The approach of Kern et al. achieved very high precision on the Small attribution tasks, but paid for it in reduced recall.

4.2 Verification

Verification results are given in Table 6. Snider et al. achieved the best precision performance overall, though not the highest recall. It should be mentioned that

Run	Macro-averaged			Micro-averaged			Rank
	Prec	Recall	F1	Prec	Recall	F1	Sum
tanguy-2011-06-07-1700	0.688	0.267	0.321	0.779	0.471	0.587	9
snider-2011-06-08-1548	0.654	0.227	0.258	0.627	0.405	0.492	20
kern-2011-06-08-1500	0.673	0.179	0.226	0.802	0.383	0.518	21
tanguy-2011-06-07-1600	0.806	0.148	0.208	0.924	0.299	0.451	26
escalante-2011-06-07-0934	0.53	0.203	0.191	0.446	0.446	0.446	29
vilarino-2011-05-31-1456	0.347	0.245	0.263	0.368	0.368	0.368	32
mikros-2011-06-08-2245	0.398	0.183	0.209	0.499	0.292	0.369	36
vilarino-2011-05-31-1455	0.488	0.084	0.088	0.222	0.222	0.222	50
ryan-2011-06-08-2331	0.19	0.154	0.132	0.216	0.216	0.216	53
eriksson-2011-06-13-0920	0.432	0.064	0.062	0.201	0.201	0.201	59
vilarino-2011-05-31-1454	0.153	0.092	0.089	0.175	0.175	0.175	63
noecker-2011-06-08-2356	0.227	0.054	0.06	0.037	0.037	0.037	70
noecker-2011-06-08-2337	0.001	0.011	0	0.001	0.001	0.001	78

Fig. 3. Results for the Large+ test set with extraneous documents.

Run	Macro-averaged			Micro-averaged			Rank
	Prec	Recall	F1	Prec	Recall	F1	Sum
tanguy-2011-06-07-1600	0.662	0.451	0.475	0.717	0.717	0.717	8
tanguy-2011-06-07-1700	0.662	0.451	0.475	0.717	0.717	0.717	8
escalante-2011-06-07-0934	0.676	0.381	0.387	0.709	0.709	0.709	19
mikros-2011-06-08-2245	0.529	0.419	0.424	0.659	0.659	0.659	28
kern-2011-06-08-1500	0.79	0.345	0.348	0.685	0.685	0.685	29
luyckx-2011-06-10-1635	0.435	0.378	0.371	0.642	0.642	0.642	39
kourtis-2011-06-08-1000	0.476	0.374	0.38	0.638	0.638	0.638	40
snider-2011-06-08-1548	0.644	0.323	0.343	0.66	0.6	0.629	45
luyckx-2011-06-10-1640	0.444	0.356	0.343	0.62	0.62	0.62	50
solorio-2011-06-08-1217	0.415	0.205	0.185	0.44	0.44	0.44	64
vilarino-2011-05-31-1456	0.236	0.284	0.358	0.432	0.432	0.432	65
vilarino-2011-05-31-1455	0.359	0.141	0.157	0.374	0.374	0.374	75
ryan-2011-06-08-2331	0.257	0.238	0.216	0.311	0.311	0.311	78
eriksson-2011-06-13-0920	0.304	0.158	0.144	0.372	0.372	0.372	80
noecker-2011-06-08-2356	0.305	0.187	0.144	0.232	0.232	0.232	84
vilarino-2011-05-31-1454	0.15	0.061	0.098	0.091	0.091	0.091	96

Fig. 4. Results for the Small test set without extraneous documents.

Run	Macro-averaged			Micro-averaged			Rank
	Prec	Recall	F1	Prec	Recall	F1	Sum
tanguy-2011-06-07-1700	0.737	0.161	0.193	0.824	0.457	0.588	14
escalante-2011-06-07-0934	0.65	0.201	0.193	0.578	0.573	0.575	16
snider-2011-06-08-1548	0.803	0.153	0.175	0.671	0.434	0.527	21
vilarino-2011-05-31-1456	0.2	0.157	0.195	0.349	0.349	0.349	31
mikros-2011-06-08-2245	0.523	0.115	0.139	0.541	0.289	0.377	34
tanguy-2011-06-07-1600	0.955	0.068	0.107	0.966	0.18	0.303	37
eriksson-2011-06-13-0920	0.462	0.087	0.077	0.331	0.331	0.331	42
kern-2011-06-08-1500	1	0.03	0.05	1	0.095	0.173	46
vilarino-2011-05-31-1455	0.371	0.077	0.084	0.301	0.301	0.301	48
ryan-2011-06-08-2331	0.209	0.137	0.101	0.254	0.254	0.254	49
noecker-2011-06-08-2356	0.275	0.086	0.06	0.189	0.189	0.189	57
vilarino-2011-05-31-1454	0.14	0.03	0.049	0.065	0.065	0.065	71

Fig. 5. Results for the Small+ test set with extraneous documents.

Run	Test set	Prec	Recall	F1	Rank Sum
snider-2011-06-08-1548	Verify1	1	0.333	0.5	9
kern-2011-06-08-1500	Verify1	1	0.333	0.5	9
vilarino-2011-05-31-1455	Verify1	0.1	0.333	0.5	12
mikros-2011-06-08-2245	Verify1	0.125	0.667	0.211	13
vilarino-2011-05-31-1456	Verify1	0.043	0.667	0.9	14
escalante-2011-06-07-0934	Verify1	0.1	0.333	0.154	17
vilarino-2011-05-31-1454	Verify1	0.033	0.333	0.5	18
tanguy-2011-06-07-1600	Verify1	0.091	0.333	0.143	20
tanguy-2011-06-07-1700	Verify1	0.091	0.333	0.143	20
eriksson-2011-06-13-0920	Verify1	0.045	0.333	0.08	24
escalante-2011-06-07-0934	Verify2	0.4	0.8	0.533	11
snider-2011-06-08-1548	Verify2	0.5	0.4	0.444	12
vilarino-2011-05-31-1455	Verify2	0.071	0.4	0.571	14
vilarino-2011-05-31-1454	Verify2	0.031	0.4	0.571	16
vilarino-2011-05-31-1456	Verify2	0.026	0.4	0.571	17
kern-2011-06-08-1500	Verify2	0.5	0.2	0.286	18
eriksson-2011-06-13-0920	Verify2	0.091	0.4	0.148	19
tanguy-2011-06-07-1600	Verify2	0.1	0.2	0.133	23
tanguy-2011-06-07-1700	Verify2	0.1	0.2	0.133	23
mikros-2011-06-08-2245	Verify2	0.035	0.6	0.067	23
snider-2011-06-08-1548	Verify3	0.211	1	0.348	9
vilarino-2011-05-31-1455	Verify3	0.091	0.333	0.5	11
vilarino-2011-05-31-1456	Verify3	0.037	0.583	0.833	12
vilarino-2011-05-31-1454	Verify3	0.034	0.333	0.5	17
tanguy-2011-06-07-1600	Verify3	0.083	0.25	0.125	17
tanguy-2011-06-07-1700	Verify3	0.083	0.25	0.125	17
eriksson-2011-06-13-0920	Verify3	0.05	0.25	0.083	21
mikros-2011-06-08-2245	Verify3	0.036	0.5	0.067	21
escalante-2011-06-07-0934	Verify3	0	0	0	30
kern-2011-06-08-1500	Verify3	0	0	0	30

Fig. 6. Results for Verification test sets (with extraneous documents).

authorship verification is considerably more difficult than authorship attribution. High precision evidently is easier to achieve than high recall.

5 Conclusions

With the great variety of feature sets and classification methods applied, it is difficult to form any overall conclusions from the basic results; more nuanced understanding will have to emerge from discussion among researchers and follow-on studies. One thing that is clear, however, is the need to decouple, to the extent possible, feature choice from classification method, so that the separate advantages and deficiencies of different feature types and algorithms can be understood, as well as their interactions. As well, one characteristic of all the better methods seems to be a preference for precision over recall (which is probably preferred in real-world applications), as in the more difficult open tasks, precision generally stayed high, while recall declined.

Regarding the different methods, the best method overall for attribution was that of Tanguy et al., who applied the largest and most diverse feature set to the problem, which may indicate the usefulness to find ways of profitably learning classifiers from very large numbers of features with diverse characters.

Acknowledgements

We would like to thank the organizers of CLEF 2011 and PAN 2011 for their support which enabled this competition to take place. Thanks are also due all the participants who made this effort such a success.

Development of this competition was funded in part by National Science Foundation grant CRI-CRD-0751198.

References

1. L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
2. W.W. Cohen and Y. Singer. A simple, fast, and effective rule learner. In *Proceedings of the National Conference on Artificial Intelligence*, pages 335–342. JOHN WILEY & SONS LTD, 1999.
3. T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.
4. N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.
5. H. Escalante, M. Montes, and L. Villaseñor. Particle swarm model selection for authorship verification. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 563–570, 2009.
6. R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.
7. P. Juola. Authorship attribution. *Foundations and Trends in information Retrieval*, 1(3):233–334, 2006.

8. M. Koppel and J. Schler. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62. ACM, 2004.
9. M. Koppel, J. Schler, and S. Argamon. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26, 2009.
10. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. *Machine Learning: ECML-98*, pages 4–15, 1998.
11. K. Luyckx and W. Daelemans. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 513–520. Association for Computational Linguistics, 2008.
12. K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67. Citeseer, 1999.
13. J.R. Quinlan. *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.
14. J.J. Rocchio. *Relevance feedback in information retrieval*. Prentice-Hall, Englewood Cliffs NJ, 1971.
15. F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.
16. E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.
17. I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, page 104. ACM, 2004.
18. H. van Halteren. Linguistic profiling for author recognition and verification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 199–es. Association for Computational Linguistics, 2004.