An Overview of the Traditional Authorship Attribution Subtask

Notebook for PAN at CLEF 2012

Patrick Juola

Evaluating Variation in Language Laboratory
Duquesne University
Pittsburgh, PA 15282
juola@mathcs.duq.edu

Juola & Associates
276 W. Schwab Ave.
Munhall, PA 15210
pjuola@juolaassoc.com

Abstract This paper describes the Traditional Authorship Attribution subtask of the PAN/CLEF 2012 workshop. As a followup to our subtask at PAN/CLEF 2011 (Amsterdam), we established a new corpus for analysis for 2012 (Rome). The new corpus differed in several ways from the previous subtask:

- Both the number and size of documents were decreased
- The documents were taken from a different genre (fiction, represented by the Feedbooks.com site)
- The documents were no longer marked up extensively
- A new sub-sub-task was added : Authorship clustering. In this new problem (related to Òintrinsic plagiarismÓ) participants were given a text of mixed authorship and asked to determine which paragraphs came from which authors.

The resulting corpus consisted of eight problems, including three closed-class authorship attribution problems, three open-class (the set of correct answers included Ònone of the aboveÓ), and two clustering problems. Twenty-five teams participated in this subtask from many different parts of the world. Detailed results are available on the Web at pan.webis.de and will be discussed in detail at the PAN/CLEF 2012 meeting in September.

1 Background

Although traditionally authorship studies are done on the basis of close reading for stylistic detail, "nontraditional" or statistical authorship attribution has been around long enough [6,1,4,5,7] to have developed into a traditional research problem of its own, especially in comparison to new tasks such as sexual predator identification [2]. The task is well-understood (given a document, determine who wrote it) although amenable to many variations (given a document, determine a profile of the author; given a document pair, determine whether they were written by the same author; given a document, determine which parts of it were written my any specific person) and the motivation is clear. Applications for this technology include not only plagiarism detection but also historical inquiry, journalism, and legal dispute resolution (forensics). TREC-style competitive analyses of authorship methods using a standardized corpus have been around since at least 2004 [3].

This competition follows on the heels of a previous subtask at the PAN 2011 conference, but differs from that competition in several ways:

1.1 Both the number and size of documents were decreased.

In last year's competition, the corpus consisted of several thousand relatively small documents, with distractor sets consisting of hundreds of authors. This was considered to create impracticalities for many participants, especially those that relied upon machine-aided instead of fully automatic analysis. We have instead focused on a smaller group of larger documents, perhaps more typical of the type of cases usually analyzed by "traditional" close reading,.

1.2 The documents were taken from a different genre.

Last year's corpus was taken from the *Enron* email corpus; this years instead was collected from the free fiction collection published by `Feedbooks.com`, including both classic fiction that is now out-of-copyright as well as (fiction, represented by the Feedbooks.com site). This of course introduces the standard issue of analysis-by-Google, but that's a very difficult problem to avoid short of generating content to order.

1.3 The documents were no longer marked up extensively.

As no one made particular use of any markup last year, the documents were simply released as text documents.

1.4 A new sub-sub-task was added : Authorship clustering.

In the most major change from last year, we created a new style of problem related to what has in prior competitions been called "intrinsic plagiarism." In this new problem participants were given a text of mixed authorship and asked to determine which paragraphs came from which authors.

2 The Problems and Corpus

2.1 Traditional authorship attribution

There were six problems of straightforward authorship attribution, presented as three pairs, representing the closed- and open- class versions of the attribution problem, respectively. In a closed-class attribution problem, a document is presented along with a set of sample authors, and the analyst or computer is asked to determine which of the sample authors wrote that document as a forced-choice scenario. In the open-class version, by contrast, "none of the above" is an acceptable answer and some of the documents to be analyzed were, in fact, written by someone other than the set of sample authors.

Problems A and B both used the same training set : two samples (each, six samples in total) by each of three authors A, B, and C. All samples were between 1800 and 6060 words. The test set for problem A consisted of six samples (two by each author); the test set for problem B consisted of a different set of six samples (two by each author) as well as four "none of the above" samples for a total of ten.

Problems C and D similarly used a shared training set, but had a larger number of authors (and hence documents). The training set had 8 authors (again, two samples per author) but were generally larger, ranging up to about 13000 words. The test set for problem C contained one sample for each training author (hence 8 documents); the test set for problem D contained one sample per in-class author, plus nine out-of-class ("none of the above") for a total of test documents.

Problems G and H were disregarded due to security issues (the test data was inadvertently released along with the training data) and replaced by problem I and J. These problems again shared a common training set, but were of novel (or at least novella) length, ranging from about 40,000 words up to about 170,000. There were 14 authors represented (the most of any task in this collection). Test data for problem I consisted of fourteen additional novels, one per candidate author; test data for problem J contained sixteen additional novels, one per candidate author plus two out-of-class novels.

The number of documents per problem approximately matches perceived difficulty of the problems; more distractor authors are more difficult, and of course open-class is more difficult than closed.

2.2 Authorship clustering

Problems E and F focused on the clustering problem; as such, no "training" data is actually needed since the point of clustering is to learn group authors based only on document-internal evidence. However, "sample" data was made available for both problems E and F to illustrate the format used.

Problem E contained intermixed paragraphs (in random order) from several different documents by different authors (one document per author; problem E1 contained two authors, E2 contained three, and E3 contained 4. Problem F by comparison contained four documents, three of which contained a single intrusive passage of several consecutive paragraphs and the final one was singly authored. All documents were segmented by paragraphs and all authorship changes occurred at paragraph boundaries. No attempt was make to control for subject or authorial voice, making this task easier than many other related plagiarism corpora.

3 Grading

Normal information retrieval measures such as precision, recall, and F-score are not really applicable to multiple (non-binary) categorization environments. Instead, documents were graded on a simple percentage-correct basis; i.e. a submission that correctly categorized four of the six documents in problem A would score 4/6 or 67% on that problem. Similarly, each paragraph in problems E/F was treated as a separate "document" and evaluated as correct or incorrect in its assignment.

Problem E was a little more complicated; each cluster identified by the participant was matched to an existing correct partition, and the number of paragraphs contained in both partitions were counted as "correct." If the number of partitions identified was incorrect, some partitions would be unmatched. Because matching can be done in many ways, we used the matching that generated the highest overall score. For example, if all

the odd numbered paragraphs were by author A, and all of the even numbers were by author B, and a participant submitted two clusters, one containing paragraphs 1–15 and another containing 16–30, there would be two possible ways of matching : odd–low (and even–high) or odd–high (and even–low). Matching odd—low generates 8 matches (paragraphs 1,3,5,7,9,11,13,15) as does even–high, for a total of 16/30 correct. Matching the other way generates 7 correct paragraphs, hence the participant would have scored 16/30 or 53% correct, the higher score.

Overall grading of a corpus like this can be slightly controversial because different approaches can yield different results on different problems. Like a decathlon (or in this case an octathlon), how does one combine different scores to an overall measure? SInce the key point of a competition such as this one is not to award medals but to encourage exploration of the field, we have taken a simple, agnostic approach to scoring and present two separate scores. The first "overall" score is the average of the individual percentages correct on all eight problems. The second "documents correct" score is the percentage of documents correctly analyzed. This second approach thus weights larger (more documents) problems more heavily, but we expect (correctly, as it turns out) that good methods will score well across all problems.

Details of the corpus are presented as table 1.

Task	Training docs	Test docs
A	6	6
B		10
C	16	8
D		17
E	n/a	90
F	n/a	80
G	n/a	n/a
H	n/a	n/a
I	28	14
J		16

Table 1. Corpus construction summarized

4 Participants and results

Twenty-five submissions were received from twelve teams. Results are summarized in table 2. A full breakdown of results including per-problem results can be obtained from the PAN website (`pan.webis.de`).

Some participants received low score due to partial submissions. For example, the Brooke submission participated in only problems E and F (authorship clustering) and did quite well on problem F in particular (41/90 for problem E, 68/80 for problem F), but did not participate in any traditional authorship problems and hence scored 0 on those.

TEAM	Overall score	Documents correct
Vilarino 1	50.45839169	59.75103734
Vilarino 2	62.13264472	63.07053942
de Graaff 1	57.54989496	21.99170124
de Graaff 2	39.47610294	15.76763485
de Graaff 3	2.941176471	1.659751037
Brainsignals	86.37429972	81.32780083
Ruseti	57.40239846	22.82157676
CLLE-ERSS 1	70.8092612	77.593361
CLLE-ERSS 2	59.12931839	68.87966805
CLLE-ERSS 3	64.70967554	64.3153527
CLEE-ERSS 4	67.66208567	73.02904564
Lip6 1	59.76759454	21.99170124
Lip6 2	54.40782563	20.33195021
Lip6 3	52.67069328	19.50207469
Bar-Ilan Univ	83.40321545	81.74273859
Sapkota	58.35346639	21.57676349
EVL Lab	81.67221055	87.96680498
Surrey	53.98663632	75.5186722
Zech terms	43.17664566	15.76763485
Zech stylo	22.91404062	8.713692946
Zech stats	30.11379552	11.2033195
Brooke	16.31944444	45.22821577
Zech I-2	17.96875	50.62240664
Zech I-3	17.03125	48.13278008
Zech I-4	16.47569444	46.47302905

Table 2. Summary results of subtask

As predicted, both scores yielded the same overall set of "winners", albeit in a different order. The top three participants were identical for both scores, as presented in table 3.

Position	Overall score	Documents correct
1st place	Brainsignals	EVL Lab
2nd place	Bar-Ilan Univ.	Bar-Ilan Univ.
3rd place	EVL Lab	Brainsignals

Table 3. Highest scoring participants

5 Future work

Assuming that future PAN participation in the CLEF experimental framework is desired (an assumption the author supports), there remain several further issues to explore.

- Both this and the previous PAN competition have focused exclusively on English documents. Should future competitions include non-English languages, and if so, which, and in what representation?
- Similarly, what genre(s) should be represented, and what sources could be used to get documents in those genres? What size of problems should be done, including both number of authors and number and size of training/test documents?
- Should we consider having documents "written to order" in order either to prevent cheating-by-Google or to ensure tighter control over the documents (especially for clustering/intrinsic plagiarism context?
- Can/should we find a way for "traditional" language scholars to participate and compare their hand analyses with the results of the computer runs?
- What other types of (sub)tasks should be attached to the competitive study of authorship attribution? Examples of potential problems might include authorship profiling (e.g. was this document written in New York, London, or California? Was it written by a man or a woman?), document dating, and so forth.
- How should future competitions be judged?

Despite the necessarily incomplete nature of any fixed-corpus study of authorship, PAN/CLEF 2012 has produced a valuable set of results and basis for discussion.

6 Acknowledgments

We want to thank all the PAN 2012 and CLEF 2012 organisers for their hard work and support, as well as the participants of the competition for their patience and suggestions. This material is based upon work supported by the National Science Foundation under Grant No. OCI–1032683. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This of course applies to the organizers and participants as well. More tersely put, the errors mine, the thanks theirs.

References

1. Holmes, D.I.: Authorship attribution. Computers and the Humanities 28(2), 87–106 (1994)
2. Inches, G., Crestani, F.: Overview of the international sexual predator identification competition at pan-2012. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers. Rome, Italy (2012)
3. Juola, P.: Ad-hoc authorship attribution competition. In: Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004). Göteborg, Sweden (June 2004)
4. Juola, P.: Authorship attribution. Foundations and Trends in Information Retrieval 1(3) (2006)
5. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology 60(1), 9–26 (2009)
6. Morton, A.Q.: Literary Detection: How to Prove Authorship and Fraud in Literature and Documents. Scribner's, New York (1978)
7. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–56 (2009)