A Trinity of Trials: Surrey's 2014 Attempts at Author Verification

Notebook for PAN at CLEF 2014

Anna Vartapetiance, Lee Gillam
University of Surrey
{A.Vartapetiance, L.Gillam}@surrey.ac.uk

Abstract. Encouraged by results from our approaches in previous PAN workshops, this paper explores three different approaches using stopword co-occurrence. High frequency patterns of co-occurrence can be used to some extent as identifiers of an author's style, and have been demonstrated to operate similarly across certain languages - without requiring deeper linguistic knowledge. However, making best use of such information remains unresolved. We compare results from applying three approaches overs such patterns: a frequency-mean-variance framework; a positional-frequency cosine comparison approach, and a cosine distance-based approach. A clearly advantageous approach across all languages and genres is yet to emerge.

1 Introduction

In the 6th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN2012), we gave first test to our ideas on co-occurrence patterns of stopwords. For three simple systems, we obtained 42.8% correct detection for Traditional Authorship Attribution, 91.1% for Intrinsic Plagiarism Detection, and 0.61, 0.38 and 0.48 for Precision, Recall and F1 respectively for Sexual Predator Identification [1]. For PAN2013, we focused only on the open class Traditional Authorship Attribution problem for three different languages (English, Greek and Spanish), and used a vector similarity approach over a frequency-mean-variance framework for patterns of a few stopwords for each language. This system achieved F1 values of 0.66, 0.74 and 0.78 for Early Bird, Final, and Post submission assessment of the Train Corpus respectively [2].

In this paper, we present 3 approaches to the PAN2014 task of Author Identification (authorship verification) involving 6 collections across 4 languages (English, Greek, Spanish and Dutch). In section 2, we briefly discuss specific changes in contrast to previous PAN efforts. In sections 3 and 4, we describe the three approaches and the evaluation of its results. Section 5 concludes the paper with considerations for future work.

2 A Brief Contrast to Previous PANs

PAN2014 focuses Authorship Attribution to a question of whether the author of a set of documents is also the author of a given document outside this set. This task covers six text collections across four different languages: English, Greek, Spanish and Dutch, and covering four genres: Essays, Reviews, Novels and Articles.

For PAN2012 [3] , given a set of documents from different known authors and a set of documents with unknown authors; the task was to allocate the documents to one author (or none). The PAN2013 [4] approach required a Boolean response as to whether an unknown document was likely written by the same author as a set of (from 1 to 10) "known" documents from that (single) author. PAN2014 now allows for three responses – introducing a "non-committal" value (0.5); better performing systems may be hedging responses rather than committing to a wrong answer. The size of the training corpus has changed substantially from 35 across three categories for PAN2013 to 696 across 6 categories for PAN2014.

Table 1, below, shows details of the six training subcorpora for PAN2014, covering numbers of cases, and averages numbers of known and unknown documents as well as statistics for the known documents. Effects of document length and number of comparisons may be useful in subsequent analysis.

Table 1: PAN2014 Corpus Details

Genre	*Case*	*Known doc. per case (max 5)*	*Avg. Unknown doc. length*	*Avg. Known doc. length*	*Avg. word per sentence -known doc.*	*Avg. of sentences -known doc.*
Dutch						
Essay	96	1.79	507.55	414.83	18.06	23.19
Review	100	1.02	122.73	125.38	18.90	6.63
English						
Essay	200	2.65	806.86	845.30	10.01	84.40
Novel	100	1.00	1783.37	4393.95	22.63	194.16
Greek						
Article	100	2.85	1447.35	1383.65	19.27	71.81
Spanish						
Article	100	5.00	1184.76	1123.23	19.85	56.58

3 Three Methods

For PAN2012, we approached attribution using a mean-variance framework on patterns of stopwords using a specified maximum window size for pairs of the 10 most common English stopwords to identify positional frequencies, and allocated an author based on nearest frequency-mean-variance match. We achieved F1 of 0.42, and saw post-submission that it might have been possible to achieve F1 of 0.48 using paired sets of 5 stopwords (i.e. patterns combined from the first 5 stopwords with the second 5, hence a smaller feature space) [1]. For PAN2013, the core approach remained the same with output adapted to the boolean output required. The task introduced Greek and Spanish texts, of which the authors have no real knowledge, and so lists of 10 frequent stopwords were sought for each.

For PAN2014, we reuse these stoplists and have now added Dutch stopwords to address Dutch subcorpora – with Dutch as yet another language of which the authors have no real knowledge.

Table 2: List of stopwords for all four languages

Language	*Stopwords*
Dutch	De Van Een Het En In Is Dat Op Te
English	The Be To Of And A In That Have I
Greek	Και Το Να Τον Η Της Με Που Την Από
Spanish	De La Que El En Y A Los Del Se

3.1 Frequency-Mean-Variance (FMV)

We follow the approach detailed at length in Vartapetiance and Gillam (2013) [2], generating frequency information for stopword pairs, determining mean and variance for separation, then applying cosine distance to compare the resulting feature vectors.

3.2 Positioning

This approach is based on FMV, above, but omits step 4 and so acts as a cosine comparison on positional frequencies for each pattern. This would tend to require comparable frequencies for each feature to ensure a good match.

3.3 Cosine

We modify the Positioning approach to consider the frequency information for all patterns as a single vector, then apply cosine distances between resulting vectors. Here we also consider how to determine a match: a single cosine distance between one known and one unknown; a difference in distance within a threshold when two known texts can be compared; and distances between the unknown and many known texts to be at a suitable point on the distribution of distances amongst knowns. Acceptability, according to thresholds, and cosine distance can then be used together to determine match confidence.

4 Submissions, Results and Evaluations

In PAN2013, we determined a set of parameters, values for which would embody language-specific treatment. In PAN2014, introduction of different text genres required additional considerations relating to the likelihood of pattern occurrence in much shorter documents. To account for such differences, we conducted a parameter sweep (over 10000 tests) based on values shown in Table below: S1, S2 and S represent first 5, second 5 and all ten most frequent stopwords respectively.

Table 3: Parameters and possible values for each

Parameter	*# of Options*	*Options*
Pattern Pairs	9	S1*S1, S1*S2, S1*S, S2*S1, S2*S2, S2*S, S*S1, S*S2, S*S
Window Size	5	5, 10, 15, 20
Filter	5	No filter, 2, 3, 4, 5
Confidence Measure	10	90, 91, 92, 93, 94, 95, 96, 97, 98, 99

Table 4 shows the values of parameters determined by this parameter sweep for FMV. Filter size can be related directly to text length in all cases except for Greek, which we attribute to the structuring.

Table 4: Values determined for the PAN2014 FMV approach

Language		*Window Size*	*Filter*	*Confidence Measure*
Dutch	Essay	5	0	0.95
	Review	10	0	0.97
English	Essay	5	4	0.92
	Novel	5	5	0.92
Greek	Article	5	3	0.97
Spanish	Article	20	5	0.99

A similar strategy leads us to the values shown in Table 5 for the Positioning and Cosine approaches.

Table 5: Values for Parameters used for PAN2014 – Positioning and Cosine Approaches

Language		*Window*	*Filter*	*Confidence*	*Measure*	*Confidence*	*Measure*
		Size		*(Positioning)*		*(Cosine)*	
Dutch	Essay	5	0		0.60		0.55
	Review	5	0		0.50		0.20
English	Essay	5	0		0.65		0.35
	Novel	5	0		0.70		0.80
Greek	Article	5	0		0.70		0.80
Spanish	Article	5	0		0.85		0.45

Table 6 shows results from the 3 approaches on 3 datasets: training, corpus-1 and corpus-2 – for all 6 categories. The best overall result is still obtained for FMV, although comparison between the values showed that the Cosine approach achieves much higher results for English Novels where the unknown documents was only being compared to 1 known document, while FMV approach had higher score for categories in which there were more known documents; e.g. Spanish with 5 known documents per test.

Table 6: Results from all three approaches for Train and Test Corpus-1 and Test Corpus-2; best results for each sub-corpus, and on average, are highlighted.

		DE	*DR*	*EE*	*EN*	*GR*	*SP*	*Average*
Training	FMV - competition	0.66	0.56	0.58	0.68	0.64	0.65	0.63
	Positioning	0.61	0.53	0.52	0.68	0.62	0.55	0.59
	Cosine	0.56	0.57	0.57	0.75	0.56	0.60	0.60
Corpus-1	FMV - competition	0.65	0.52	0.55	0.52	0.54	0.60	0.56
	Positioning	0.48	0.48	0.54	0.50	0.58	0.60	0.53
	Cosine	0.50	0.56	0.60	0.56	0.60	0.56	0.56
Corpus-2	FMV - competition	0.72	0.51	0.52	0.50	0.53	0.66	0.57
	Positioning	0.56	0.57	0.46	0.53	0.49	0.58	0.53
	Cosine	0.54	0.56	0.57	0.56	0.51	0.55	0.55

5 Conclusion

In this paper, we attempted to reuse and adapt a fairly simple approach from PAN2013 for Authorship Attribution. Our frequency-mean-variance framework demonstrates reasonable performance (0.63) on training data, and similar (0.57) on test data. Our positioning approach is less performative (0.59 and 0.53), and cosine approach sits between these two (0.60 and 0.55). These results suggest that a broader grain in comparison achieves a marginally better result than a positional focus might offer, which indicates that the direction of future exploration needs to account for this broader grain.

Acknowledgments

The authors gratefully acknowledge prior from EPSRC/JISC (EP/I034408/1), the UK's Technology Strategy Board (TSB, 169201), and also the efforts of the PAN organizers in crafting and managing the tasks.

References

[1] A. Vartapetiance and L. Gillam, "Quite Simple Approaches for Authorship Attribution , Intrinsic Plagiarism Detection and Sexual Predator Identification -Notebook for PAN at CLEF 2012," in *Working Notes Papers of the CLEF 2012 Evaluation Labs*, 2012.

[2] A. Vartapetiance and L. Gillam, "A Textual Modus Operandi: Surrey's Simple System for Author Identification - Notebook for PAN at CLEF 2013," in *Working Notes Papers of the CLEF 2013 Evaluation Labs*, 2013.

[3] P. Juola, “An Overview of the Traditional Authorship Attribution Subtask - Notebook for PAN at CLEF 2012,” in *Working Notes Papers of the CLEF 2012 Evaluation Labs*, 2012.

[4] P. Juola and E. Stamatatos, “Overview of the Author Identification Task at PAN 2013,” in *Working Notes Papers of the CLEF 2013 Evaluation Labs*, 2013.