A Textual Modus Operandi: Surrey's Simple System for Author Identification

Notebook for PAN at CLEF 2013

Anna Vartapetiance, Lee Gillam

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

Abstract. Detecting deceptions of various kinds may be variously possible, but has little value if the deceiver cannot be identified. In this paper, we discuss our approach to Authorship Attribution that uses vector similarity with a frequency-mean-variance framework for patterns of stopwords (no more than ten). The high frequency individual occurrences, and patterns of co-occurrence, can be used as identifier of an author's style, and operates similarly across certain languages without prior linguistic knowledge. This simple system achieved F1 values of 0.66, 0.74 and 0.78 for Early Bird, Final, and Post submission assessment of the Train Corpus. We cannot yet offer further explanation as the Test Corpus is not available at the time of writing.

1 Introduction

Research into Deception Detection has benefited from the large (documented) sets of human communication mediated through the web and in particular through social media. Asynchronous distributed communication is common in such media, and with the non-verbal and vocal cues to deception removed, as well as the deceiver having time to plan their deception, verbal cues are the main area of exploration. Such detection is attempted on simple text messages [7], fraud investigations [6] and court testimonies [4]. Deceptions range from "Pareto white lies" to "Spite black lies" [2], and include studies by forensic linguists and natural language processors alike. Detecting the deception differs, however, from detecting the deceiver – analogous to the difference between analysing the scene of a crime and being able to use specific evidence from that scene to suggest the perpetrator of the crime. Extending the analogy, we are interested in a detectable *Modus Operandi* (MO) for a particular perpetrator. However in the PAN problem space of Authorship Attribution, we are trying to denote whether a given 'scene' or 'design' reflects the MO of (a) prior scene(s) or design(s).

In the 6th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN2012), we gave first test to our ideas that the signatures of such scenes could be found in co-occurrence patterns of stopwords. PAN2012's task covered [10]:

1) Traditional Authorship Attribution: given unknown documents and sets of known documents from different authors, the task was,

    a) to denote an author for each document (closed class problem)
    b) an extension to a) where the author may have been somebody else (open class problem)
2) Authorship clustering/intrinsic plagiarism: given a document,
    a) Clustering the paragraphs written by each author – where the number of authors are known (closed class clustering)
    b) Clustering the paragraphs written by each author – where the number of authors are unknown (open class clustering)
3) Sexual Predator Identification, given a datasets of chat lines,
    a) identify whether the chat indicated a predator
    b) identify the predatory elements of the chat

We submitted simple systems for all three subtasks to create baselines for our own work. The results achieved 42.8% of overall correct detection for Traditional Authorship Attribution, 91.1% for Intrinsic Plagiarism Detection. For Sexual Predators Identification tasks, our system achieved 0.61, 0.38 and 0.48 for Precision, Recall and F1 respectively.

This paper, presents our approach for PAN2013 focusing only on the open class Traditional Authorship Attribution problem for three different languages (English, Greek and Spanish). The approach, the dataset, and the addition of two languages are significant changes, making it inherently difficult to infer performance from prior results and so making it likewise difficult to determine whether a given approach adapted to this task offers better or worse performance without incurring a cost of back-fitting.

In this paper, we outline the approach taken at the University of Surrey to this task. In section 2, we discuss the Train Corpus and highlight the changes compared to last year. In sections 3 and 4, we describe the system and the evaluation of its results. Section 5 concludes the paper with considerations for future work.

2 Corpus

PAN2013 focuses on an open class Traditional Authorship Attribution for three different languages -English, Greek and Spanish. PAN2012 had a related task, but any prior approach could not be used directly, and the addition of languages likely requires further adaptation.

For PAN2012, 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 approach requires a Boolean response as to whether an unknown document was likely written by the same author as a set of (1 to 10) “Known Documents” from that (single) author. Table 1, shows details of the corpus with the number of cases per each language and the number of “Known Documents” available per each case to help predicting the correct answer.

As is apparent from the table, there are neither an equal number of cases provided for each language nor identical number of “Known Documents” for each. Nor is it clear that the numbers are representative of those used for the Test Corpus.

Table 1: PAN2013 Corpus Details

Language	*Cases*	*# of Known Documents per case*	*Case Names*
English	10	2	EN11, EN21, EN23, EN30
		3	EN13, EN18,
		4	EN07
		5	EN24
Greek	20	1	GR01, GR02
		2	GR03, GR04
		3	GR05, GR06
		4	GR07, GR08
		5	GR09, GR10
		6	GR11, GR12
		7	GR13, GR14
		8	GR15, GR16
		9	GR17, GR18
		10	GR19, GR20
Spanish	5	1	SP03, SP09
		3	SP02, SP05
		4	SP10

3 Method

For previous Authorship Attribution tasks, many approaches have been documented that use NLP techniques over bags of words, N-grams, and parts of speech (POS), with varying degrees of success. Often, stopwords are either not an integral part of the analysis, or are dropped from processing. For PAN2012, we approached attribution using a mean-variance framework on patterns of stopwords [1]. We used 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 match 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 (e.g. patterns combined from the first 5 with the second 5, and hence a smaller feature space) [10].

For PAN2013, this core idea was not changed. The authors have no real knowledge of either Greek or Spanish, so attempted to find lists of 10 frequent stopwords for each (Table 2). Given that lack of linguistic knowledge, we do not yet know whether the lists we obtained meet this requirement.

Table 2: List of stopwords for all three languages

Language	*Stopwords*	*Based on*
English	The Be To Of And A In That Have I	[9]
Greek	Και Το Να Τον Η Της Με Που Την Από	[8]
Spanish	De La Que El En Y A Los Del Se	[3]

For PAN2013 early bird submission, we applied the following steps with parameters from our PAN2012 post-submission experiments. Patterns were generated from the first 5 frequent stopwords against the second 5 frequent stopwords, with window size of 5 words, and confidence measure of 0.95. We replaced our closest match option from PAN2012 with the average of maximum cosine similarity values per pattern. The approach was:

Table 3: Approach taken for PAN2013 Early Bird Submission

Steps	***Process***
Step 1	Select the 10 most frequent words for each language
Step 2	Generate regular expressions of first 5 most frequent stopwords against the second 5 (S1*S2) and use a specific size of window N (here, 5) for each document
Step 3	Extract concordances containing the regular expressions for all texts
Step 4	Calculate frequency, mean and variance information for the pairs
Step 5	Calculate cosine similarities of the unknown document against each of the known documents per pair
Step 6	Calculate the average of all maximum cosine similarities for pairs to get a single value per case
Step 7	Report “Y” if the value is above the confidence measure (here, 0.95), “N” otherwise

For the main submission, we introduced a filter (after Step 4) to only compare patterns that exist more than a specified number of cases in one document. For example, just one occurrence of a pattern may not a strong indicator for an author’s writing style.

An algorithm of the system, using the denotations and functions from Table 4 is offered in Table 5.

Table 4: Table of Notations

Symbol	***Meaning***
Q	Set of Queries
q	A single query where $q \in Q$
D	Set of documents
d	A document where $\{d_{01}, d_{02}, \dots, d_N\} \in D$
D_q	Set of documents D related to query q
L	Set of languages
sw	A Stopword
S_L	Set of stopwords ($sw_{L,1}$, $sw_{L,2}$ … $sw_{L,H}$) for a language L
S_a, S_b	Subsets of S_L,where
	$S_a = S_b \in (S_1 \vert S_2 \vert S_3) \Rightarrow \begin{cases} S_1 = \{S_i \vert\, 1 \leq i \leq \lceil {}^1/_2\ length_{(S_L)} \rceil\} \\ S_2 = \{S_j \vert \lceil {}^1/_2\ length_{(S_L)} \rceil + 1 \leq j \leq length_{(S_L)}\} \\ S_3 = S_L \end{cases}$
WS	Window Size: maximum distance from S_a to S_b ,where $WS \in \mathbb{N}$

$\mathrm{PP}^{ws}\ (\mathrm{X},\mathrm{Y})$	Pattern of stopword X from S_a followed by Y from S_b in maximum distance of Window Size WS
FT	Filter: threshold for frequency of each pattern, where $FT \in \mathbb{N}$
CM	Confidence Measure: threshold for identifying confidence in similarity of Q with D, where $CM \in \{1,2,3,\dots,99,100\}$
FMV	Function that takes the incidents of given pattern $\mathrm{PP}^{ws}\ (\mathrm{X},\mathrm{Y})$ and returns three values of frequency, mean, and variance
CosineSim	Cosine Similarities function [5] where $\cos(A \cdot B) = \frac{A \cdot B}{\lvert A\rvert\ \lvert B\rvert}$

Table 5: Algorithm of our System for PAN2013

```
Algorithm
$\boldsymbol{for}\ all\ q\ \boldsymbol{do}$
    $\boldsymbol{for}\ all\ X \leftarrow 1\ to\ length\ S_a\ \boldsymbol{and}\ all\ Y \leftarrow 1\ to\ length\ S_b\ \boldsymbol{do}$
        $Sum_q(X,Y) = 0$
        $\boldsymbol{for}\ ws \leftarrow 0\ to\ WS\ \boldsymbol{do}$
            $\boldsymbol{if}\ \mathrm{PP}_q^{ws}\ (X,Y)\ \boldsymbol{then}$
                $Count_q\ [ws](X,Y) += 1$
                $Sum_q(X,Y) += 1$
        $\boldsymbol{if}\ Sum_q(X,Y) \geq FT\ \boldsymbol{thens}$
            $FMV_q(X,Y) \leftarrow FMV_q\left(Count_q\ [ws](X,Y)\right)$
            $\boldsymbol{for}\ all\ D_q\ \boldsymbol{do}$
                $Sum'_d(X,Y) = 0$
                $\boldsymbol{for}\ ws \leftarrow 0\ to\ WS\ \boldsymbol{do}$
                    $\boldsymbol{if}\ \mathrm{PP}_d^{ws}\ (X,Y)\ \boldsymbol{then}$
                        $Count'_d\ [ws](X,Y) += 1$
                        $Sum'_d(X,Y) += 1$
                $\boldsymbol{if}\ Sum'_d(X,Y) \geq FT\ \boldsymbol{then}$
                    $FMV_d(X,Y) \leftarrow FMV_d(Count'_d\ [ws](X,Y))$
                    $CosineSim_q(X,Y) \leftarrow$
                    $CosineSim_{q,\,D_q}\left(FMV_q(X,Y), FMV_{D_q}(X,Y)\right)$
        $MaxCosineSim_q(X,Y) \leftarrow Max\ \left(CosineSim_q(X,Y)\right)$
    $\boldsymbol{if}\ MaxCosineSim_q(X,Y) \neq 0\ \boldsymbol{then}$
        $RES_q \leftarrow AVG\ (MaxCosineSim_q(X,Y))$
$\boldsymbol{if}\ RES_q \geq CM\ \boldsymbol{return}$
    $"Match"$
$\boldsymbol{else\ return}$
    $"No\ Match"$
```

Our process of Authorship Attribution can be explained as:

1. For all the $q \in Q$, calculate the FMV with pair of X from Pattern set S_a followed by Y from Pattern set S_b within window size of WS; only if pattern has happened more than FT times

2. Only for Patterns that happened more that FT times for q, for related D_q calculate the FMV with pair of X from Pattern set S_a followed by Y from Pattern set S_b within window size of WS if that pattern has happened more than FT times too
3. Find maximum of Cosine similarities ($MaxCosineSim$) between each of the patterns for q and related D_q
4. Calculate average of non-zero $MaxCosineSim$ values
5. Answer "$Match$" if that value is bigger than Confidence Measure CM, else answer "$No\ Match$"

4 Submissions, Results and Evaluations

For early bird evaluation, we used the same parameters for all three languages following the steps presented in Table 2 (using (S1*S2) pattern in a Window Size of 5 and Confidence Measure of 95). The system achieved F1 of 0.66 for the Train Corpus, detecting 60%, 60% and 100% of documents correctly for English, Greek and Spanish respectively (Table 8). The results for first evaluation on the Test Corpus showed F1 of 0.56, detecting 45%, 50% and 90% for English, Greek and Spanish respectively.

To try to improve results, we conducted a parameter sweep that covered 6750 tests based on the values outlined below.

Table 6: Presenting Parameters and Options used for each

Parameter	*# of Options*	*Options*
Language	3	English, Greek, Spanish
Pattern Pairs	9	S1*S1, S1*S2, S1*S3, S2*S1, S2*S2, S2*S3, S3*S1, S3*S2, S3*S3
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

The results from these tests suggested that each language should be treated slightly differently. Although we do not have linguistic knowledge of Greek or Spanish, Greek seemed to evidence more structured use of stopwords than Spanish (high cosine similarities for Greek suggested stopwords occupy relatively fixed positions which makes them less author specific than would be the case for Spanish). For the full submission, parameters were selected for each language – to account for these findings - as follows:

Table 7: Values for Parameters used for PAN2013 Final submission

Language	*Pattern Pairs*	*Window Size*	*Filter*	*Confidence Measure*
English	S1*S2	20	4	92
Greek	S3*S3	10	5	98
Spanish	S1*S2	10	4	92

These parameters improved performance of our Early Bird system from F1 of 0.66 to F1 of 0.74 (presented in Table 8). However, results from Test Corpus on Final Submission showed F1 of only 0.54 across the three languages, a significant difference (Spanish dropped by F1 of 0.30, while both English and Greek improved). Unfortunately, the Test Corpus has not been released at the time of writing, and so we are unable to offer an explanation of this variation.

Post-competition submission, we could indicatively achieve F1 of 0.78 on the Train Corpus by considering a factor of the number of test samples (Known Documents) being compared against. The value of this finding would need to be explored once all test data and suitable annotations become available.

Table 8: Results from Various Submission for both Train and Test Corpus

Version	*E*	*G*	*S*	*E%*	*G%*	*S%*	*Overall*	*Corr doc*	*F1*
Train 1	6	12	5	60	60	100	73.3	23	0.657
Test- Early Bird	--	--	--	45	50	90	61.6	--	0.56
Train 2	8	13	5	80	65	100	81.6	26	0.742
Test- Final Sub	--	--	--	50	53	60	53.3	--	0.541
Train- Post sub	8	15	5	80	75	100	85	28	0.777

5 Conclusion

In this paper, we attempted to reuse a fairly simple approach from PAN2012 for Authorship Attribution. Our frequency-mean-variance framework over pairs of stopwords (no more than ten) can demonstrate reasonable performance F1 of 0.74 on Train Corpus, but seems only to achieve F1 of 0.54 on Test Corpus suggesting either that our approach is overturned to training data, or that we suffer from generalizability problems (not having more similar data to test with to tune parameters) or that there is a big gap in representatively between Train and Test Corpus. Only once these data are released could we ascertain which.

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 PAN13 organizers in crafting and managing the tasks.

References

1. Church, K., Hanks, P.: Word Association Norms, Mutual Information and Lexicography. Computational Linguistics, vol. 16(1), pp. 22-29 (1991)

2. Erat, S., Gneezy, U.: White lies. Journal of Management Science, vol. 58 (4), pp. 723-733 (2012)
3. Lazarinis, F.: Engineering and Utilizing a Stopword List in Greek Web Retrieval. Journal of the American Society for Information Science and Technology (JASIST), vol. 58(11), pp. 1645-1652 (2007)
4. Little, A., Skillicorn, B.: Detecting Deception in Testimony. In: Proceeding of IEEE International Conference of Intelligence and Security Informatics (ISI 2008), pp. 13-18. Taipei, Taiwan (2008).
5. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, USA (2008)
6. McCallion, J.: Ernst & Young Debuts FBI Co-designed Anti-Fraud Software. In: IT PRO. [available at] http://www.itpro.co.uk/644899/ernst--young-debuts-fbi-co-designed-anti-fraud-software
7. Reynolds, L., Smith, M.E., Birnholtz, J., Hancock, J.: Butler Lies From Both Sides: Actions and Perceptions of Unavailability Management in Texting. In: Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13), pp. 769-778. ACM, New York, NY, USA (2013)
8. Snowball: A Spanish stop word list. [available at] http://snowball.tartarus.org/algorithms/spanish/stop.txt
9. The Oxford English Corpus: Facts about the language. [available at] http://oxforddictionaries.com/words/the-oec-facts-about-the-language
10. Vartapetiance, A., Gillam, L.: Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification - notebook for pan at clef 2012. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.): CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers. Rome, Italy (2012)