UNIVERSITY OF SURREY

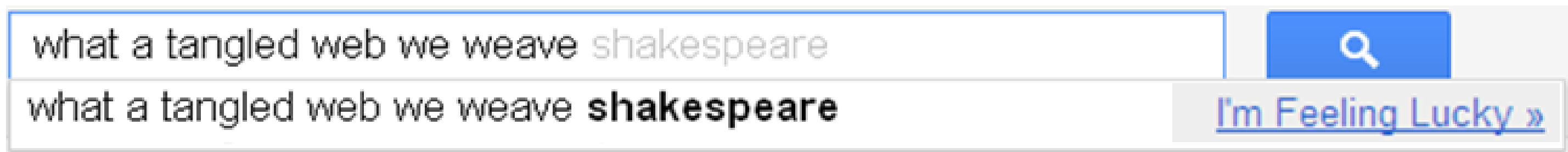
A Textual Modus Operandi: Surrey's System for Author Identification

Notebook for PAN at CLEF 2013

Anna Vartapetiance and Dr. Lee Gillam
{A.Vartapetiance, L.Gillam}@surrey.ac.uk

Introduction

If we simply let machines learn, will humans end up being deceived? What Google would suggest for an author of this particular phrase may not coincide with reality. Correct authorship attribution is but one part of our deception detection research.



what a tangled web we weave shakespeare

what a tangled web we weave **shakespeare**

Aims and Objectives

PAN2013 has an open class Traditional Authorship Attribution task. Given an "Unknown Document " and a (set of) "Known Document" from a single author (in three different languages of English, Greek and Spanish) identify:

a) Yes – the same author
b) No – not the same author

Method

In PAN2012 [1], we used a frequency-mean-variance framework over patterns of stopwords [2] achieving f1 of 0.42 in the open class part of the test corpus with potential for f1 of 0.48 (post-submission analysis).

For PAN2013 [3] we are using cosine distances over this frequency-mean-variance framework.

~~Bag of words N gram Part of Speech SVM Machine Learning~~

Stopwords	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

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 $(sw_{L,1}, sw_{L,2} \dots sw_{L,H})$ for a language L
S_a, S_b	Subsets of S_L,where $S_a = S_b \in (S_1 \mid S_2 \mid S_3) \Rightarrow \begin{cases} S_1 = \{S_i \mid 1 \le i \le \lceil {}^1\!/_2\ length_{(S_L)} \rceil\} \\ S_2 = \{S_j \mid \lceil {}^1\!/_2\ length_{(S_L)} \rceil + 1 \le j \le length_{(S_L)}\} \\ S_3 = S_L \end{cases}$
WS	Window Size: maximum distance from S_a to S_b ,where $WS \in \mathbb{N}$
$\mathrm{PP}^{\mathrm{ws}}$ (X, 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 $\mathrm{PP}^{\mathrm{ws}}$ (X, Y) and returns three values of frequency, mean, and variance
CosineSim	Cosine Similarities function [5] where $\cos(A \cdot B) = \frac{A \cdot B}{\lvert A \rvert\ \lvert B \rvert}$

Defining the Approach

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 "$No\ Match$"

Algorithm

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

Results and Evaluation

We conducted a parameter sweep that covered 6750 tests based on the values outlined below

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

Parameters chosen for the final submission based on the highest scores where:

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

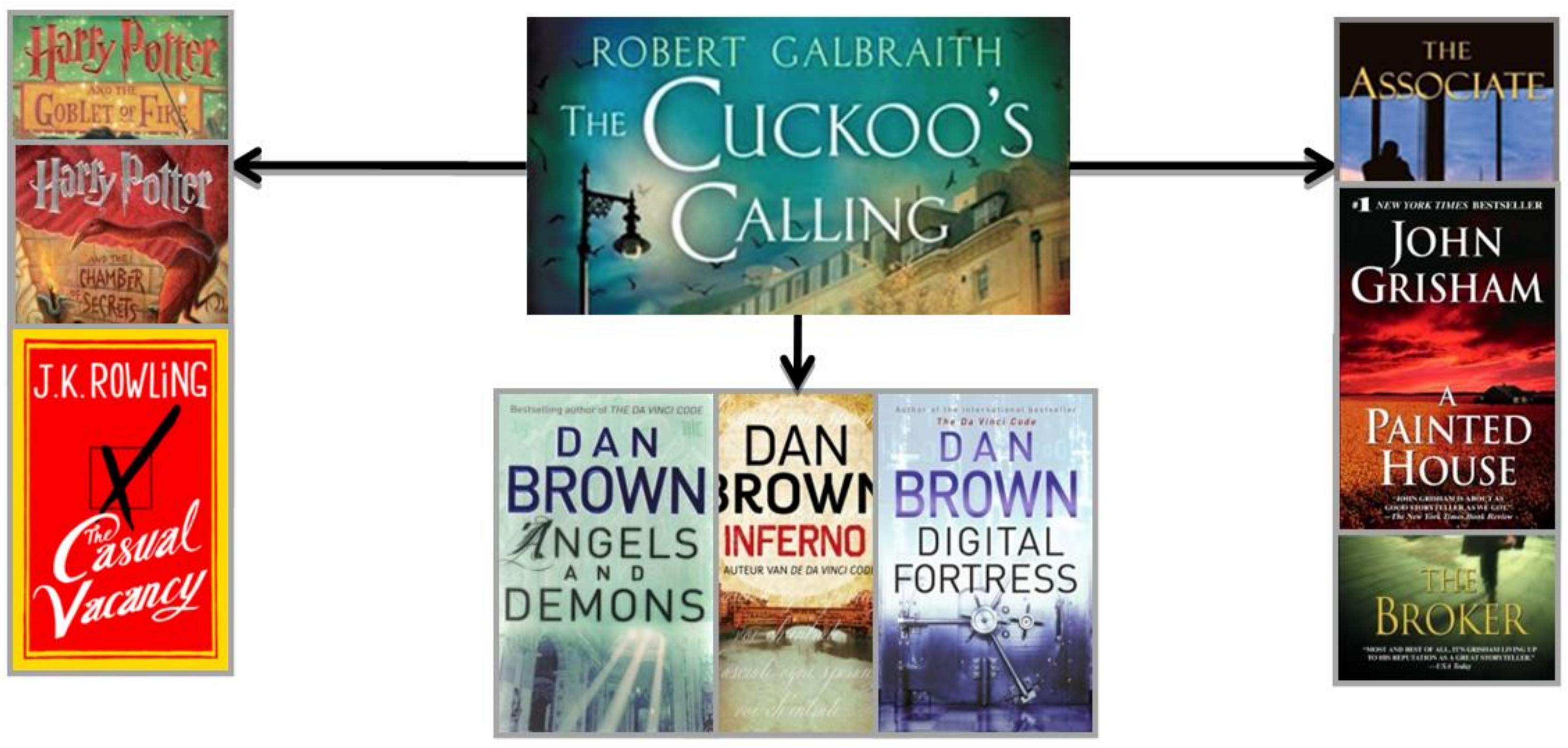
Table below shows the results from different experiments on Train and Test datasets[**Note:** The test data has not yet been released, hence, surprising decline in the final results for Spanish language can not yet been explained]

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**

Conclusion

Our frequency-mean-variance framework over pairs of stopwords (no more than ten) can demonstrate reasonable performance (f1 of 0.74 on training corpus). Post-submission experiments improve slightly (0.78) by considering the number of known files an unknown documents is compared to (e.g. more or less that 5)

Deception and Authorship Attribution



Authors' Unique Pattern in Using Stopwords

The Cuckoo's calling	There, flinging discretion to the chilly wind (in a most un-Matthewlike way), he had proposed, on one knee, in front of three down-and-outs huddled on the steps, sharing what looked like a bottle of meths. It had been, in Robin's view, the most perfect proposal, ever, in the history of matrimony. He had even had a ring in his pocket, which she was now wearing; a sapphire with two diamonds, it fitted perfectly, and all the way into town she kept staring at it on her hand as it rested on her lap. She and Matthew had a story to tell now, a funny family story, the kind you told your children, in which his planning (she loved that he had planned it) went awry, and turned into something spontaneous.
The Casual Vacancy	He had endured a thumping headache for most of the weekend and was struggling to make a deadline for the local newspaper. However, his wife had been a little stiff and uncommunicative over lunch, and Barry deduced that his anniversary card had not mitigated the crime of shutting himself away in the study all morning. It did not help that he had been writing about Krystal, whom Mary disliked, although she pretended otherwise. Mary had softened and smiled, so Barry had telephoned the golf club, because it was nearby and they were sure of getting a table. He tried to give his wife pleasure in little ways, because he had come to realize, after nearly two decades together, how often he disappointed her in the big things.
Angels and Demons	After passing through endless security checks and being issued a six-hour, holographic guest pass, he was escorted to a plush research facility where he was told he would spend the afternoon providing "blind support" to the Cryptography Division; an elite group of mathematical brainiacs known as the code-breakers. For the first hour, the cryptographers seemed unaware Becker was even there. They hovered around an enormous table and spoke a language Becker had never heard. They spoke of stream ciphers, self-decimated generators, knapsack variants, zero knowledge protocols, unicity points. Becker observed, lost. They scrawled symbols on graph paper, pored over computer printouts, and continuously referred to the jumble of text on the overhead projector.
The Associate	Another dumb foul and Kyle yelled at the referee to just let it slide. He sat down and ran his finger over the side of his neck, then flicked off the perspiration. It was early February, and the gym was, as always, quite chilly. Why was he sweating? The agent/cop hadn't moved an inch; in fact he seemed to enjoy staring at Kyle. The decrepit old horn finally squawked. The game was mercifully over. One team cheered, and one team really didn't care. Both lined up for the obligatory high fives and "Good game, good game," as meaningless to twelve-year-olds as it is to college players. As Kyle congratulated the opposing coach, he glanced down the court. The white man was gone. What were the odds he was waiting outside?

Cosine Similarity based on Patterns of Stopwords

Unknown	Author	Books in the Corpus	Cosine Value
The Cuckoo's Calling	J.K. Rowling	The Sorcerer's Stone, The Chamber of Secrets, The Prisoner of Azkaban, The Goblet of Fire, The Deathly Hallows, The Casual Vacancy	99.92
	Dan Brown	Digital Fortress, Inferno, Angles and Demons	99.54
	John Grisham	The Appeal, The Innocent Man, The Associate, Bleachers, A Painted House, The Broker	99.43

References

[1] 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)

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

[3] Vartapetiance, A., Gillam, L.:A Textual Modus Operandi: Surrey's System for Author Identification - notebook for pan at clef 2013. In: P. Forner, R. Navigli, and D. Tufis (eds). CLEF 2013 Evaluation Labs and Workshop –Working Notes Papers, Valencia, Spain (2013)

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.