Adapting for Subject-Specific Term Length using Topic Cost in Author Verification

Notebook for PAN at CLEF 2015

Anna Vartapetiance and Lee Gillam

Department of Computing, University of Surrey, UK
{a.vartapetiance, l.gillam}@surrey.ac.uk

Abstract. Previous PAN workshops have offered us the opportunity to explore three different approaches using basic statistics of stopword pairs for author verification. In this PAN, we were able to select our 'best' approach and explore the question of how authors writing about different subjects would necessarily adapt to term lengths specific to the subject. The adaptation required is, essentially, a redistribution of frequency: where longer terms occur. We introduce the notion of a 'topic cost' which increases the propensity for matching. Results show AUC and C1 scores of 0.51, 0.46 and 0.59 for Dutch, Greek and Spanish respectively. The English results are not yet available, as the evaluation system was unable to run the approach due to as yet unknown reasons.

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 [1]. At the 8th iteration (PAN 2014), we presented 3 variations to our approach, largely geared around evaluating use of similarity/distance over vector spaces [2].

In this paper, we suggest extension to our approaches to the PAN2014 by accounting for a 'topic cost'. Simply, there are several reasons why specific stopword-pair separation may be less able to indicate similarity, and accounting for term length and term count offers potential for addressing this. In section 2, we briefly discuss the previous approaches we have used for author verification. Section 3 explains how we determine and use topic cost. Section 4 offers results and evaluation, and Section 5 concludes the paper.

2 Previous methods applied

As discussed in [1], for PAN2012, we approached author 'attribution' using a mean-variance framework on patterns of stopwords with 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.

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 reused these stoplists along with a stoplist for Dutch – with Dutch as yet another language of which the authors have no real knowledge. We also evaluated 3 approaches based on:

Frequency-Mean-Variance: We follow the approach detailed at length in Vartapetiance and Gillam 2013, generating frequency information for stopword pairs, determining mean and variance for separation, then applying cosine distance to compare the resulting feature vectors.

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.

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.

3 PAN 2015

For this year's task, we wanted to explore the ability to match where the same author may necessarily vary their writing according to the topic. This would account for, say, simple temporal modification– discussing for example 'the former Prime Minister of' rather than 'the Prime Minister of' – but is principally geared to account for differences in term lengths as relate to topics. In the 'Prime Minister' example given, the same stopword pair of the-of is present, but with a positional mismatch. Since position, and variability in position, is core to our approaches, we require a simple way to address the pattern-specific positional mis-alignment that occurs.

To approach this, we introduce the notion of a 'topic cost' and distribute positional frequencies according to this topic cost. To determine topic cost, we simply count the number of terms and the length of these terms, and use the difference between these values for redistribution. The only additional resource employed is a language-specific stoplist as exposes the terms.

As an example, consider the following passage of text:

> *UK interest rates have been kept unchanged again by the Bank of England, meaning they have now been at their record low of 0.5% for six years. Rates*

were first cut to 0.5% in March 2009 as the Bank sought to lift economic growth amid the credit crunch.

Take stopword pairs as formed from [*the, of, in, for, to*]**.** If we ignore the sentence break, the first pair of interest offers us: "*for six years. Rates were first cut to*". The distance covered by the pair is 6 (the number of words between "for" and "to"). Collecting all multi-word terms, using all stopwords (not just those listed) as delimiters (and, here, the full-stop also), results in 3 terms comprising 5 words – six years, rates, first cut. The topic cost, then, is 2. Instead of counting once at position 6, we uniformly distribute – other weightings possible but unexplored - across position 6 and the two preceding positions and so positions 4, 5 and 6 each receive 0.333. This example, and further from the above passage, are shown in the table below.

Table 1: Example of 'Topic Cost' applied on sample sentence

Extracted text	***Gap***	***Remove all stops***	***Topic cost***	***Shift (word, gap, count)***
for six years. Rates were first cut to	6	six years rates first cut	2	*for-to, 6, 1 becomes* for-to, 6, 0.333 for-to, 5, 0.333 for-to, 4, 0.333
to 0.5% in	1	0.5%	0	No change
to lift economic growth amid the	4	lift economic growth amid	3	*to-the, 4, 1 becomes* to-the, 4, 0.25 to-the, 3, 0.25 to-the, 2, 0.25 to-the, 1, 0.25
in March 2009 as the	3	March 2009	1	*in-the, 3, 1 becomes* in-the, 3, 0.5 in-the, 2, 0.5
the Bank of	1	Bank	0	No change
the Bank sought to	2	Bank sought	1	*the-to, 2, 1 becomes* the-to, 1, 0.5 the-to, 2, 0.5

In principle, use of topic cost offers greater potential for match using our previous approaches. In practice, the extent of improvement over previous results is likely to be marginal.

4 Results

Results for each of the PAN 2015 collections are shown in the table below based on 4 language categories.

Table 2: **Results from our approaches for Test Corpus**

Collection	***AUC***	***C1***	***Score***
Dutch	0.51	0.51	0.262
English	---	---	---
Greek	0.46	0.46	0.212
Spanish	0.59	0.59	0.348

Due to yet unknown problem with English run, the system was unable to calculate the outcomes of the test. Also, unfortunately, the results from the runs using last year's systems will not be available until after this paper is submitted, so the authors are not able to provide a comparison between systems to see whether or not this approach improves the outcome of detection. However, the results on runs on training datasets using FMV, Positioning and Topic Cost systems (Table 3) show some improvements in detection using the new system.

Table 3: Results from FMV, Positioning and Topic Cost systems based on Training Corpus

Collection	***AUC***		
	FMV	***Positioning***	***Topic Cost***
Dutch	**0.5**	0.49	0.46
English	0.46	0.51	**0.53**
Greek	0.45	0.51	**0.56**
Spanish	0.54	0.55	**0.56**
Average	0.49	0.52	**0.53**

5 Conclusions and Future Work

In this paper, we suggested an extension to our approaches to PAN2014 for authorship verification by accounting for a 'topic cost'. For us, topic cost may account for lower match values in our previous approaches, and our intention was to determine whether a simple treatment of topic cost could improve our results. This modification does require much more testing in respect to the test collections of previous years to fully appreciate its effect. Unfortunately, other activities hindered the authors' abilities to allocate sufficient time to this testing during this round of PAN.

Acknowledgements

The authors gratefully acknowledge prior funding from 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 Trinity of Trials : Surrey ’ s 2014 Attempts at Author Verification Notebook for PAN at CLEF 2014,” *Work. Notes Pap. CLEF 2014 Eval. Labs*, 2014.