A simple Local n-gram Ensemble for Authorship Verification

Notebook for PAN at CLEF 2014

Robert Layton

Internet Commerce Security Laboratory
Federation University Australia r.layton@icsl.com.au

Abstract The authorship verification task requires deciding whether a given test document was written by the same author as a training set. For my attempt I tested a simple voting ensemble of local (character) n-gram methods, using a grid search to choose parameters. This results in a method that requires little pre-configuration and can be applied to any language with a concept of characters. The method itself is quite fast, however training is slow with the large number of attempted parameter combinations. The approach results in accuracies of around 60% depending on the corpus and application.

1 Introduction

Local n-gram (LNG) methods employ a profile based character n-gram approach to authorship analysis. An author's profile consists of the top L n-grams for a given author, different to many feature selection methods which usually use a globally relevant set of features. The author profile is then compared to a similarly created document profile using a given distance metric. For authorship attribution, a classification task, each candidate author is first profiled. A test document of unknown authorship is then profiled and the distance to each candidate author profile is computed. The author with the lowest distance is predicted as the author.

Applying LNG to authorship verification makes use of the concept of distance, but applies in a different way. First we formally define the problem.

We are given a training set of documents D (usually such that $1 <= |D| <= 5$) all authored by the same person A. Next, we are given a test document d_t. The task is to determine whether the author of d_t is A, i.e. the author of the documents in D. We refer to the above task as a single trial, with the authorship verification task composing of a large number of trials in different languages and different contexts.

We employed a straight-forward translation of the use of LNG for classification purposes to authorship verification purposes. We calculated the distance between the test document d_t and each of the documents in the training set D, called the inter-distance, and compared that to the internal distance between documents in D to themselves, the intra-distance. The assumption was that if the inter-distance was approximately equal to the intra-distance, then the document was likely to be from the same author. If they were not approximately equal, then it is more likely a different author wrote d_t.

1.1 Summary of Results

Our results were quite poor for this year's competition. The focus on automation of the algorithm may have hurt performance, due to a lack of tuning. It would be recommended to use a more fine-tuned approach, rather than a generic 'catch-all' approach.

2 Datasets

There were six datasets in the training corpus, and a total of 596 individual trials. There were four languages represented in the released datasets; Dutch, English, Greek and Spanish. Compositions of the datasets are provided in table 1, and all datasets had approximately equal number of positive and negative trials.

Language	Context	Trials	Positive:Negative	Documents per Trial	Characters per Document
Dutch	Essays	96	47:49	1.79	4,342
Dutch	Reviews	100	50:50	1.02	689
English	Essays	200	100:100	2.65	12,683
English	Novels	100	50:50	1.00	25,402
Greek	News	100	50:50	2.85	26,761
Spanish	News	100	50:50	5.00	34,443

Table 1. Statistics of the six datasets

3 Local n-gram Methods

In recent years, the authorship analysis field has used machine learning techniques for a majority of research [15]. In this paper, we too focus on such techniques. Algorithms in authorship analysis can be placed into two categories; global and local methods. Global methods fit a more standard feature based machine learning methodology. In this methodology, a set of features is used to take measurements of each of a set of documents. This gives us a matrix X such that $X_{i,j}$ is the value of feature j of document i. This model can be used as input into a large number of classification or clustering algorithms, such as Support Vector Machines or the k-means algorithm.

Advances in local algorithms have shown great success in this alternate form of model. A document, or set of documents, is represented as a profile P such that $P(x)$ is the value of feature x for the document, or set of documents. The features chosen are usually character n-grams, subsequences of continuous characters or length n. The value of $P(x)$ is then given as the frequency of n-gram x in the document, or set of documents, being profiled. When local models are used with character n-grams, the approach is called Local n-grams (LNG). What makes a local model particularly different from a global model is that there is no global set of features. By profiling the set of all documents known to be from one author, we can profile that author's writings. While most applications of LNG have been supervised, LNG methods have also been used for unsupervised methodologies, outperforming feature based models [13].

Instead, each document (or set) is profiled using the set of features *most distinctive* to that document (or set). This means each profile has a separate set of features associated with it. The phrase *most distinctive* usually means 'most frequent', however the RLP algorithm introduced later has a different definition.

As notation for the following, we state that a feature x is 'in' a profile P if $P(x) \neq 0$. The intersection of two profiles is the set of features that are in both, and the union is the set of features in either (ignoring values).

The first model of this type was the Common n-grams (CNG) method [7]. A profile is given as the set of the L most frequent n-grams, for some value of L. Profiles are then compared using equation 1. A document of unknown authorship is attributed to the author with the most similar profile.

$$K(P_1, P_2) = \sum_{x \in X_{P_1} \cup X_{P_2}} \left(\frac{2 \cdot (P_1(x) - P_2(x))}{P_1(x) + P_2(x)} \right)^2 \tag{1}$$

A variant of this form, the Source Code Author Profile (SCAP) algorithm, was introduced by [6]. This algorithm is a variant of CNG, with only one change. Rather than using equation 1 to compare profiles, the similarity of two profiles is given as the size of the set intersection of them. The higher the number of features in the intersection, the more similar they are. This is bounded by the choice of L as an input, and therefore can be normalised by dividing by L. The major finding by [6] was that this approach approximated the results of [7] using a much simpler algorithm. SCAP can be very fast to run on modern systems and provides a good approximation to CNG, allowing it's use in prototyping [10]. CNG-WPI (Weighted Profile Intersection) is an improvement to SCAP which weights the n-grams based on the number of documents they appear in (inferring the likelihood of the n-gram to appear in both profiles) [5].

Stamatatos' $d1$ and $d2$ measures are improvements designed to work with unbalanced datasets [14]. Their approaches weighted the profile similarity comparison using a profile of language default values. This approach was found to be more effective for imbalanced dataset than CNG, while less effective for balanced dataset.

The Recentred Local Profiles (RLP) algorithm was developed by [11], again using this concept of a language default profile. The derivation of the CNG methodology used work by [1], which originally included a concept of a language default value; the expected frequency of a particular n-gram in normal use of the language. This language default was removed in a simplification of the algorithm which lead to the CNG methodology. RLP reinstated this component, which adjusted profile weights based on this language default value, such that $P_D(x) = P_D^C(x) - P_L(x)$, where P_D^C is the LNG profile of the doucment, P_L is the CNG profile of all documents in that language. A profile then consisted of the L most distinctive n-grams, i.e. those with the highest absolute weights. Documents are then compared using a variant of the cosine distance metric, given in equation 2.

$$R(P_1, P_2) = 1 - \frac{P_1 \cdot P_2}{||P_1||_2 ||P_2||_2} \tag{2}$$

LNG methods have shown a high accuracy in difficult domains [9, 8, 14, 4]. In addition, little recoding is needed to apply them in multiple languages. Almost every

language has a definition of 'character', and for those that do not, LNG methods can be applied to byte level n-grams, rather than characters [6, 3]. These methods are remarkably resilient to the language, and can be applied to source code languages, as well as native languages [2, 12].

4 LNG Methods Employed

For this application, we used the CNG, SCAP and RLP methods as base level classifications. These methods were chosen as a representative sample of LNG methods, not specifically due to the relative efficacy, as other LNG methods achieve similar accuracies. The n values chosen for the task were 3 to 5 inclusive, while L values of 1000, 2000, 5000 and 10000. This set of parameters was chosen to be small enough to compute in a reasonable time, while still being relatively representative of feature values proven effective in other studies.

For each of the based methods (CNG, SCAP and RLP), a grid search of parameters was conducted to find the most accurate. The grid search compared all combinations of n and L values

5 Applying Thresholds for Verification

We used a distance based threshold for determining whether the given document belonged to a specific author. This threshold was relative to the documents themselves, not a global value.

We calculated the distance between the test document d_t and each of the documents in the training set D, called the inter-distance, and compared that to the internal distance between documents in D to themselves, the intra-distance. The assumption was that if the inter-distance was approximately equal to the intra-distance, then the document was likely to be from the same author. If they were not approximately equal, then it is more likely a different author wrote d_t.

Because there was only one test document, we could not apply standard statistical distribution comparisons, and instead opted for simpler approach. The inter-distance was considered to be approximately equal to the intra distance if it was less than the average intra-distance, plus two standard deviations (of the intra-distance for a given dataset D).

The obtained results were less than expected, and less than expected for the individual based models. This suggests that this threshold method may require extensive work, and perhaps an alternate strategy.

6 Summary of Results

While there was some variability to the results and rank (with a top rank of 4th on one corpus), the results were typically quite poor. The baseline for the results is approximately 0.5, which was beaten in all datasets, but only barely in most. The performance was also less than expected, based on cross-validation results within the dataset. The

Language	Context	AUC	C1	Score	Rank
Dutch	Essays	0.54557	0.5625	0.30688	12/13
Dutch	Reviews	0.5026	0.5200	0.26135	10/13
English	Essays	0.5947	0.6100	0.36277	4/13
English	Novels	0.51	0.51	0.2601	12/13
Greek	News	0.6612	0.6100	0.40333	8/13
Spanish	News	0.5534	0.5400	0.29884	12/13

Table 2. Results of the six datasets

reason for this is likely the small amount of data was not properly accounted for in the cross-validation model, meaning that the final model overfit the data. The early investigations into these errors suggests that this is caused by the threshold based method, and not the baseline LNG methods. There were some errors that can be traced to the baseline LNG methods though, suggesting that further testing is necessary for this form of application.

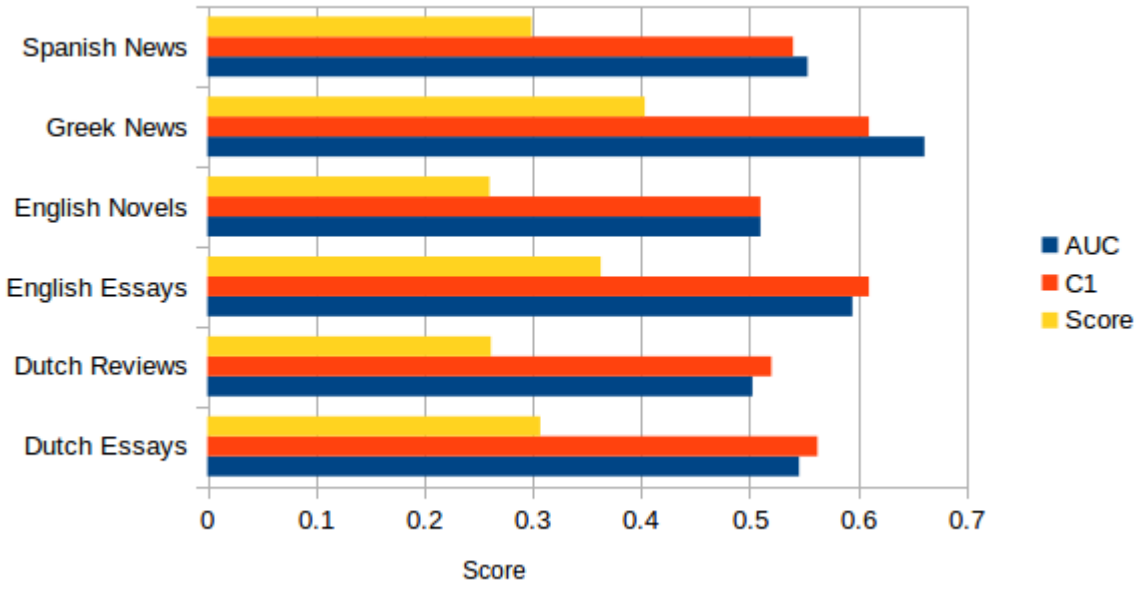

Figure 1. Results on each dataset

Bibliography

[1] Bennett, W.R.: Scientific and engineering problem-solving with the computer. Prentice Hall PTR Upper Saddle River, NJ, USA (1976)

[2] Burrows, S., Tahaghoghi, S.M.: Source code authorship attribution using n-grams. In: Proceedings of the Twelth Australasian Document Computing Symposium, Melbourne, Australia, RMIT University. pp. 32–39 (2007)

[3] Burrows, S., Uitdenbogerd, A.L., Turpin, A.: Application of information retrieval techniques for source code authorship attribution. In: Database Systems for Advanced Applications. p. 699âĂŞ713 (2009)

[4] Chatzicharalampous, E., Frantzeskou, G., Stamatatos, E.: Author identification in imbalanced sets of source code samples. In: Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on. vol. 1, p. 790âĂŞ797 (2012)

[5] Escalante, H.J., Montes-y Gómez, M., Solorio, T.: A weighted profile intersection measure for profile-based authorship attribution. In: Advances in Artificial Intelligence, pp. 232–243. Springer (2011)

[6] Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C.E.: Identifying authorship by byte-level n-grams: The source code author profile (SCAP) method. Int. Journal of Digital Evidence 6 (2007)

[7] Kešelj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING. vol. 3, p. 255âĂŞ264 (2003)

[8] Layton, R., McCombie, S., Watters, P.: Authorship attribution of irc messages using inverse author frequency. In: Cybercrime and Trustworthy Computing Workshop (CTC), 2012 Third. pp. 7–13. IEEE (2012)

[9] Layton, R., Watters, P., Dazeley, R.: Authorship attribution for twitter in 140 characters or less. In: Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second. pp. 1–8. IEEE (2010)

[10] Layton, R., Watters, P., Dazeley, R.: Automatically determining phishing campaigns using the uscap methodology. In: eCrime Researchers Summit (eCrime), 2010. pp. 1–8. IEEE (2011)

[11] Layton, R., Watters, P., Dazeley, R.: Recentred local profiles for authorship attribution. Natural Language Engineering 18(03), 293–312 (2012)

[12] Layton, R., Watters, P., Dazeley, R.: Unsupervised authorship analysis of phishing webpages. In: Communications and Information Technologies (ISCIT), 2012 International Symposium on. pp. 1104–1109. IEEE (2012)

[13] Layton, R., Watters, P., Dazeley, R.: Automated unsupervised authorship analysis using evidence accumulation clustering. Natural Language Engineering 19(01), 95–120 (2013)

[14] Stamatatos, E.: Author identification using imbalanced and limited training texts. In: Database and Expert Systems Applications, 2007. DEXA'07. 18th International Workshop on. p. 237âĂŞ241 (2007)

[15] Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3), 538–556 (2009)