UniNE at CLEF 2016: Author Clustering
Notebook for PAN at CLEF 2016

Mirco Kocher
University of Neuchâtel
rue Emile Argand 11
2000 Neuchâtel, Switzerland
Mirco.Kocher@unine.ch

Abstract. This paper describes and evaluates an effective unsupervised author clustering authorship linking model called SPATIUM-L1. The suggested strategy can be adapted without any problem to different languages (such as Dutch, English, and Greek) in different genres (*e.g.*, newspaper articles and reviews). As features, we suggest using the *m* most frequent terms of each text (isolated words and punctuation symbols with *m* at most 200). Applying a simple distance measure, we determine whether there is enough indication that two texts were written by the same author. The evaluations are based on six test collections (PAN AUTHOR CLUSTERING task at CLEF 2016).

1 Introduction

The authorship attribution problem is an interesting problem in computational linguistics but also in applied areas such as criminal investigation and historical studies where knowing the author of a document (such as a ransom note) may be able to save lives. With the Web 2.0 technologies, the number of anonymous or pseudonymous texts is increasing and in many cases one person writes in different places about different topics (*e.g.*, multiple blog posts written by the same author). Therefore, proposing an effective algorithm to the authorship problem presents a real interest. In this case, the system must regroup all texts by the same author (written according to different genres) into the same group or cluster. A justification supporting the proposed answer and a probability that the given answer is correct can be given to improve the confidence attached to the response (Savoy, 2016).

This author clustering task is more demanding than the classical authorship attribution problem. Given a document collection the task is to group documents written by the same author such that each cluster corresponds to a different author. The number of distinct authors whose documents are included is not given. This task can also be viewed as establishing authorship links between documents and is related to the PAN 2015 task of authorship verification.

This paper is organized as follows. The next section presents the test collections and the evaluation methodology used in the experiments. The third section explains our proposed algorithm called SPATIUM-L1. In the last section, we evaluate the proposed scheme and compare it to the best performing schemes using six different test collections. A conclusion draws the main findings of this study.

2 Test Collections and Evaluation Methodology

To evaluate the effectiveness of a clustering algorithm, the number of tests must be large and run on a common test set. To create such benchmarks, and to promote studies in this domain, the PAN CLEF evaluation campaign was launched (Stamatatos *et al*., 2016). Multiple research groups with different backgrounds from around the world have participated in the PAN CLEF 2016 campaign. Each team has proposed a clustering strategy that has been evaluated using the same methodology. The evaluation was performed using the *TIRA* platform, which is an automated tool for deployment and evaluation of the software (Gollub *et al.*, 2012). The data access is restricted such that during a software run the system is encapsulated and thus ensuring that there is no data leakage back to the task participants (Potthast *et al.*, 2014). This evaluation procedure also offers a fair evaluation of the time needed to produce an answer.

During the PAN CLEF 2016 evaluation campaign, six collections were built each containing six problems (training + testing). In each problem, all the texts matched the same language, are in the same genre, and are single-authored, but they may differ in text-length and can be cross-topic. The number of distinct authors is not given. In this context, a problem is defined as:

> *Given a collection of up to 100 documents, identify authorship links and groups of documents by the same author.*

The six collections are a combination of one of three languages (English, Dutch, or Greek) and one of two genres (newspaper articles or reviews). An overview of these collections is depicted in Table 1. The training set will be used to evaluate our approach and the test set will be used in order to be able to compare our results with those of the PAN CLEF 2016 campaign.

Table 1. PAN CLEF 2016 *training* corpora statistics

		Training Sets		
Corpus	Texts	Authors	Single	Words
English Newspaper	50	35; 25; 43	27; 17; 37	741; 745; 734
English Reviews	80	55; 70; 40	39; 62; 17	969; 1080; 1020
Dutch Newspaper	57	51; 28; 40	46; 20; 32	1,086; 1,334; 1,026
Dutch Reviews	100	54; 67; 91	31; 44; 83	128; 135; 126
Greek Newspaper	55	28; 38; 48	10; 26; 42	756; 750; 735
Greek Reviews	55	50; 28; 40	46; 13; 29	534; 646; 756

For each benchmark we have three problems in the training dataset containing the same number of texts with the exact corresponding number given under the label "Texts". The number of distinct authors for each problem is indicated in the column "Authors", and the number of authors with only a single document under the label "Single". For example, with the English newspaper collection (training set), 50 texts are written by 35 authors and in this text subset we can find 27 authors who wrote only one single article. These metrics are not available for the test corpora because the

datasets remained undisclosed thanks to the *TIRA* system. We only know that the same combinations of language and genre are present.

When inspecting the training collection of Dutch reviews, the number of words available is rather small (in mean 130 words for each document). Overall, there are many authors who only wrote a single text, so the number of authors per problem is rather large. This means we should only cluster two documents if there are enough signs for a single authorship.

During the PAN CLEF 2016 campaign, a system must return two outputs in a JSON structure. First, the detected groups have to be written to a file indicating the author clustering. Each document has to belong to exactly one cluster, thus the clusters have to be non-overlapping. Second, a list of document pairs with a probability of having the same author has to be written to another file representing the authorship links.

As performance measure, two evaluation measures were used during the PAN CLEF campaign. The first performance measure is the BCubed F-Score (Amigo *et al.*, 2007) to evaluate the clustering output. This value is the harmonic mean of the precision and recall associated to each document. The document precision represents how many documents in the same cluster are written by the same author. Symmetrically, the recall associated to one document represents how many documents from that author appear in its cluster.

As another measure, the PAN CLEF campaign adopts the mean average precision (MAP) measure for the authorship links between document pairs (Manning *et al.*, 2008). This evaluation measure provides a single-figure measure of quality across recall levels. The MAP is roughly the average area under the precision-recall curve for a set of problems. Therefore, this measure gives more emphasis on the first positions and a misclassification with a lower probability is less penalized.

Considering the six benchmarks as a whole, we have 18 problems to solve and 18 problems to train (pre-evaluate) our system. Because there are many authors with only a single document, we can compare our approach with a naïve baseline, which clusters each text in an individual cluster. This means the document precision is always 100%. The documents recall is lower, but should still be competitive due to the low number of expected clusters. Furthermore, random scores are assigned for all combinations in the authorship links.

3 Simple Clustering Algorithm

To solve the clustering problem, we suggest an unsupervised approach based on a simple feature extraction and distance measure called SPATIUM-L1 (Latin word meaning distance). The selected stylistic features correspond to the top *m* most frequent terms (isolated words without stemming but with the punctuation symbols). For determining the value of *m*, previous studies have shown that a value between 200 and 300 tends to provide the best performance (Burrows, 2002; Savoy 2016). Some documents were rather short and we further excluded the words only appearing once in the text. This filtering decision was taken to prevent overfitting to single occurrences. The effective number of terms *m* was set to at most 200 terms but was in most cases well below. With this reduced number the justification of the decision will be simpler

to understand because it will be based on words instead of letters, bigrams of letters or combinations of several representation schemes or distance measures.

To measure the distance between one document A and another text B, SPATIUM-L1 uses the L1-norm as follows:

$$\Delta(A,B) = \Delta_{AB} = \sum_{i=1}^{m} |P_A[t_i] - P_B[t_i]| \quad (1)$$

where m indicates the number of terms (words or punctuation symbols), and $P_A[t_i]$ and $P_B[t_i]$ represent the estimated occurrence probability of the term t_i in the first text A and in the other text B respectively. To estimate these probabilities, we divide the term occurrence frequency (tf_i) by the length in tokens of the corresponding text (n), Prob[t_i] = tf_i / n, without smoothing and therefore accepting a 0.0 probability.

To verify whether the resulting Δ_{AB} value is small or rather large, we need to have a comparison. To achieve this, the distance from A to all other k texts from the current problem was calculated. If this Δ_{AB} value is 2.0 standard deviations below the average of all distances, then this is a first indication of an author link. Since the m terms are always selected from the first text, the Δ_{AB} value might be different from the Δ_{BA} value. We therefore calculate the distance of text B with all other k texts and if this Δ_{BA} value is as well 2.0 standard deviations below the average of all distances, then this is our second indication of an author link. The exact difference to the mean divided by the standard deviation is used to calculate how much the indication weights, where a higher number means more evidence of a shared authorship. For example, in the second English review problem we cluster document 9 together with document 50. The $\Delta_{9;50}$ value is 32, while the average $\Delta_{9;B}$ value to all other texts is 45 with a standard deviation of 5.6, which results in a first indication of $(45 - 32)/5.6 = 2.3$. A higher value means more evidence of a shared authorship.

For the grouping stage we follow the transitivity rule. If we have enough indication that the texts A and B are written by the same author and we also have indication that the documents B and C have a single authorship, then we will group A, B, and C together even if we don't have enough evidence that A and C have the same writer.

For the author link, on the other hand, we only report A-B and B-C as a having the same author in this scenario, while leaving out A-C due to the absence of any previous sign for a single authorship. Furthermore, since this step allows a ranked listing of the author links, we assigned the highest probability to the text pair where we have the most evidence. A rather low probability is attributed to document pairs where we only have partial indication of a shared authorship.

4 Evaluation

Since our system is based on an unsupervised approach we were able to directly evaluate it using the training set. In Table 2, we have reported the same performance measure applied during the PAN CLEF campaign, namely the BCubed F-Score and the MAP. Each collection consists of three sets of problems and we report the average of them. The final score is the mean between the two reported metrics.

Table 2. Evaluation for the six *training* collections

Corpus	**Final**	F-Score	MAP
English Newspaper	**0.4116**	0.7915	0.0317
English Review	**0.4144**	0.8036	0.0252
Dutch Newspaper	**0.4720**	0.8230	0.1210
Dutch Review	**0.4285**	0.8201	0.0369
Greek Newspaper	**0.4804**	0.8239	0.1368
Greek Review	**0.5590**	0.8480	0.2700
Overall	**0.4610**	0.8184	0.1036
Naïve Baseline	**0.4169**	0.8115	0.0222

.

The algorithm returns the best results for the Greek Review collection with a final score of 0.5590 followed by the Greek Article and Dutch Article corpora. The worst result is achieved with the two English collections which are slightly worse than the Dutch Review corpus. For the two Dutch collections we can clearly see the difference in text length reflected in the final score, as the newspaper corpus contains almost 10 times more words and achieves a noteworthy higher value. Our approach achieves an F-Score that is slightly higher than the one from the naïve baseline, but a significantly higher MAP.

The test set is then used to rank the performance of all 7 participants in this task. Based on the same evaluation methodology, we achieve the results depicted in Table 3 corresponding to the six test corpora.

As we can see, the final score with the Greek Review corpus is the highest as expected from the training set. The results we achieved in the two English collection is as low as in the training set. On the other hand, the Greek result achieved for the newspaper part is only slightly worse than the estimation from the training set. Generally, we see a very similar performance when comparing it with the training set. Therefore, the system seems to perform stable independent of the underlying text collection and is not over-fitted to the data.

Table 3. Evaluation for the six *test* collections.

Corpus	**Final**	F-Score	MAP
English Newspaper	**0.4295**	0.8159	0.0431
English Review	**0.4207**	0.8199	0.0214
Dutch Newspaper	**0.4291**	0.8160	0.0421
Dutch Review	**0.4302**	0.8135	0.0468
Greek Newspaper	**0.4370**	0.8191	0.0548
Greek Review	**0.4814**	0.8467	0.1160
Overall	**0.4379**	0.8218	0.0540
Naïve Baseline	**0.4187**	0.8209	0.0165

To put those values in perspective we can see in Table 4 our result in comparison with the top three of all participants using macro-averaging for the effectiveness measures and showing the total runtime. We have also added our naïve baseline as

described above. As in the training collections, our approach achieves an F-Score that is slightly higher than the one from the naïve baseline, but a significantly higher MAP. Therefore, some documents were wrongly clustered together, which decreases the document precision part of the BCubed F-Score. But we cluster many documents correctly together (increases document recall) and assign them a high score for their authorship link (increases MAP). Overall, this is beneficial and we are ranked second out of eight approaches.

Table 4. Evaluation comparison.

Rank	User	**Final**	F-Score	MAP	Runtime (h:m:s)
1	bagnall16	**0.4956**	0.8223	0.1689	63:03:59
2	kocher16	**0.4379**	0.8218	0.0540	00:01:50
3	Naïve Baseline	**0.4187**	0.8209	0.0165	00:00:34
4	sari16	**0.4176**	0.7952	0.0399	00:07:48
…	…	**…**	…	…	…

The runtime only shows the actual time spent to classify the test set. On *TIRA* there was the possibility to first train the system using the training set which had no influence on the final runtime. Since we have an unsupervised system it did not need to train any parameters, but this possibility might have been used by other participants. Overall, we achieve excellent results using a rather simple and fast approach in comparison with the other solutions[1].

In text categorization studies, we are convinced that a deeper analysis of the evaluation results is important to obtain a better understanding of the advantages and drawbacks of a suggested scheme. By just focusing on overall performance measures, we only observe a general behavior or trend without being able to acquire a better explanation of the proposed assignment. To achieve this deeper understanding, we could analyze some problems extracted from the English corpus. Usually, the relative frequency (or probability) differences with very frequent words such as *when*, *is*, *in*, *that*, *to*, or *it* can explain the decision.

5 Conclusion

This paper proposes a simple unsupervised technique to solve the author clustering problem. As features to discriminate between the proposed author and different candidates, we propose using at most the top 200 most frequent terms (words and punctuations). This choice was found effective for other related tasks such as authorship attribution (Burrows, 2002). Moreover, compared to various feature selection strategies used in text categorization (Sebastiani, 2002), the most frequent terms tend to select the most discriminative features when applied to stylistic studies (Savoy, 2015). In order to take the author linking decision, we propose using a simple distance measure called SPATIUM-L1 based on the L1 norm.

[1] http://www.tira.io/task/author-clustering/

The proposed approach tends to perform very well in three different languages (Dutch, English, and Greek) and in two genres (newspaper articles and reviews, but keeping the same genre inside a given test collection). Such a classifier strategy can be described as having a high bias but a low variance (Hastie *et al.*, 2009). Changing the training data does not change a lot the decision. However, the suggested approach ignores other significant information such as mean sentence length, POS (part of speech) distribution, or topical terms. Even if the proposed system cannot capture all possible stylistic features (bias), changing the available data does not modify significantly the overall performance (variance).

It is common to fix some parameters (such as time period, size, genre, or length of the data) to minimize the possible source of variation in the corpus. However, our goal was to present a simple and unsupervised approach without many predefined arguments.

With SPATIUM-L1 the proposed clustering could be clearly explained because it is based on a reduced set of features on the one hand and, on the other, those features are words or punctuation symbols. Thus the interpretation for the final user is clearer than when working with a huge number of features, when dealing with *n*-grams of letters or when combing several similarity measures. The SPATIUM-L1 decision can be explained by large differences in relative frequencies of frequent words, usually corresponding to functional terms.

To improve the current classifier, we will investigate the consequence of some smoothing techniques, the effect of other distance measures, and different feature selection strategies. In the latter case, we want to maintain a reduced number of terms. In a better feature selection scheme, we can take account of the underlying text genre, as for example, the most frequent use of personal pronouns in narrative texts. As another possible improvement, we can ignore specific topical terms or character names appearing frequently in an author profile, and terms that can be selected in the feature set without being useful in discriminating between authors.

Acknowledgments

The author wants to thank the task coordinators for their valuable effort to promote test collections in author clustering. This research was supported, in part, by the NSF under Grant #200021_149665/1.

References

1. Amigo, E., Gonzalo, J., Artiles, J., & Verdejo, F. 2009. A comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints. *Information Retrieval*, 12(4), 461-486.
2. Burrows, J.F. 2002. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3), 267-287.
3. Gollub, T., Stein, B., & Burrows, T. 2012. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Hersh,

B., Callan, J., Maarek, Y., & Sanderson, M. (eds.) SIGIR. *The 35th International ACM*, 1125–1126.

4. Hastie, T., Tibshirani, R., & Friedman, J. 2009. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer-Verlag: New York (NY).
5. Manning, C.D., Raghaven, P., & Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
6. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., & Stein, B. 2014. Improving the Reproducibility of PAN's Shared Tasks: - Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Handbury, A., & Toms, E. (eds.) CLEF. *Lecture Notes in Computer Science*, vol. 8685, 268–299. Springer.
7. Savoy, J. 2016. Estimating the Probability of an Authorship Attribution. *Journal of American Society for Information Science & Technology*, 67(6), 1462-1472.
8. Savoy, J. 2015. Comparative Evaluation of Term Selection Functions for Authorship Attribution. *Digital Scholarship in the Humanities*, 30(2), 246-261.
9. Sebastiani, F. 2002. Machine Learning in Automatic Text Categorization. *ACM Computing Survey*, 34(1), 1-27.
10. Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M. 2016. Clustering by Authorship Within and Across Documents. In Working Notes Papers of the CLEF 2016 Evaluation Labs, *CEUR Workshop Proceedings*, CEUR-WS.org.