A Basic Character N-gram Approach to Authorship Verification

Notebook for PAN at CLEF 2013

Michiel van Dam

Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Netherlands
m.c.vandam@student.tudelft.nl

Abstract This paper describes our approach to the Author Identification task in the PAN 2013 evaluation lab. We use a profile-based approach and use the common n-grams (CNG) method that employs a normalized distance measure for short and unbalanced text introduced by Stamatatos[6]. We achieved the 9th place with an overall F_1 score of 0.6.

1 Introduction

Textual plagiarism can be described as "The unacknowledged use of an exact copy or a slightly modified version of the text written by another author." Automated methods to detect plagiarism can approach this in two ways: either by looking at the textual content to detect passages that align within two documents, or by looking at the writing style of an author and detecting changes in the style when a plagiarized passage is encountered. The first approach is suitable for extrinsic analysis, when a large reference corpus is present from which plagiarized passages are suspected to originate. The latter is more suitable for intrinsic analysis when no reference corpus is available and plagiarism can be found by detecting writing style changes within a single document.

This paper describes our approach to the Author Identification subtask, that was sumbitted to the PAN 2013 evaluation lab, and is structured as follows. First in section 2 the submitted algorithm is explained in detail, then in section 3 the results of this algorithm for the Author Identification subtask are discussed, ending in section 4 with some final notes.

2 Author Identification Task

The PAN 2013 Author Identification task focuses on determining whether an unknown document has the same author as a given set of known documents that are all written by a single author. For the task corpus the documents within one problem are matched for genre, register, theme and date of writing, making the problem closely resemble an intrinsic plagiarism analysis question: are there significant differences in writing style between two sections of the same document?

Our algorithm works as follows: first the texts are pre-processed, to increase the accuracy of the author profiles. Then author profiles are created for the set of known documents and for the unknown document, based on the Common N-Grams (CNG) approach[2]. Thirdly the distance between the known author profile and the unknown author profile is calculated, and finally based on the reported distances a judgement is made whether the unknown author is the same person as the known author.

Each of these steps is discussed in more detail in the next subsections, including the algorithm configuration for English and Greek. For Spanish it was decided to use the same approach as for English, rather than evaluate choices on the small provided training set. For English the training data was expanded using training instances generated from books in Project Gutenberg[3]. An overview of the algorithm configuration for each language is given in table 1.

Feature	**English & Spanish**	**Greek**
n-gram length	4	3
Profile length	2300	1500
Preprocessing	digit replacement no punctuation no capitalization no diacritics	digit replacement no punctuation no capitalization

Table 1. An overview of the algorithm configuration.

2.1 Text pre-processing

It was already presented in [1] that simple text transformations can yield considerable improvements in accuracy. For that reason we adopted the suggested digit transformation where all digits are replaced by a special symbol '@', because the important stylistic information is the use of digits rather than the exact combination of digits.

Secondly because the limited amount of text for every test instance, being on average 1000 words for every document, we felt the need to reduce redundancy in character n-grams when words occur both capitalized at the beginning of a sentence and lowercase in the middle of a sentence, and common verb conjugations occurring both right before a comma and a space etc. We chose to remove all punctuation except spaces, and to convert each document to lowercase for this purpose. While this will discard stylistic information contained in capital and punctuation use, we expect this to strengthen the stylistic features concerning word preferences and conjugation usage.

With the same reasons for removing capitalization, we decided to remove diacritics for English and Spanish. After experiments on the training corpus it was determined that removing diacritics lowered accuracy for Greek, which can be explained by the polytonic orthography used in Greek, where authors and publishers sometimes still use diacritics heavily for indicating accents.

2.2 Common n-grams approach

The CNG method was introduced by [2] and is a language-independent approach that has given good results for many authorship questions. The CNG method represents each document as a bag of character *n*-grams and is a profile-based approach, meaning that the known texts for an author are concatenated and the resulting large text is used for extracting the author profile. The author profile is the list of character *n*-grams with their frequency of occurrance, normalized for the length of the text. For the unknown document a similar profile is extracted.

In [5] it was already concluded that for the CNG method *n*-grams of length $3 <= n <= 5$ and a profile length L of $1000 <= L <= 5000$ usually gives the best results. Furthermore from [2] can be concluded that 3 is a good choice for n when processing Greek documents.

Experiments done on the training corpus confirmed 3 as a good n-gram size for Greek and suggested 4-grams for English. Experiments on varying the profile length indicated 2300 and 1500 as a good profile length for 4-grams and 3-grams respectively. It often happens that the ordered n-grams from $L-x$ to $L+y$ have the same normalized frequency, and in such cases we chose to also include all n-grams with the same size as the Lth n-gram, making 1500 and 2300 the *minimum* profile lengths.

2.3 Calculating judgements

Having extracted an author profile $P(A)$ for the unknown document and a profile $P(B)$ for the set of known author documents, the distance between these profiles is calculated using the normalized distance function nd_1 proposed by [6]. This distance function was proposed for comparing the writing style in one section of text to the style of a whole document, making it appropriate for this author identification task.

$$nd_1(A,B) = \frac{\sum_{g \in P(A)} \left(\frac{2(f_A(g) - f_B(g))}{f_A(g) + f_B(g)} \right)^2}{4|P(A)|}$$

To provide easy judgements on whether a specific distance $d_{A,B}$ indicates the same or a different author we make use of the corpus characteristics: within the corpus there are an equal amount of cases with the same author and different authors. Therefor within a single language the average distance $\bar{d}$ is calculated from all reported distances, and a distance $d_{A,B} < \bar{d}$ is taken to indicate that documents A and B have the same author.

3 Evaluation

As a first-time contestant and despite using a simple approach we achieved an overall 9th place (out of 17) in the Author Identification subtask of the PAN 2013 evaluation labs. Our algorithm had a runtime of roughly 9 seconds, making it the 3rd fastest algorithm. Table 2 shows the detailed performance for our contribution.

The difference in performance for English and Spanish is remarkable, especially because the accuracy for English is much lower than that for Spanish, and the algorithm

hasn't been designed with Spanish in mind. A possible explanation for these results is in the pre-processing steps that are taken, which can reduce the available stylistic information in different amounts for different languages.

The very low performance for Greek, which is around the 50% that is attainable by random guess, is unexplained as of yet. While the research in [2] shows a good performance for character n-gram techniques on Greek, those results were attained for a closed class attribution problem using roughly 10 times the available data for this task, and a possible explanation could be sought in the differences between the experiments performed there and in this PAN 2013 subtask.

Language	**F1**	**Precision**	**Recall**
English	0.600	0.600	0.600
Greek	0.467	0.467	0.467
Spanish	0.760	0.760	0.760

Table 2. The attained performance on the PAN 2013 test data.

4 Conclusion and Future Work

The presented algorithm performs above the baseline for English and Spanish, but is failing for Greek. A good direction for further work would be an instance-based approach on the available training texts where more than one text is provided for an author, to get a better estimate of the average distance between documents of that author. This estimate could be used instead of the global average distance for all authors of a language.

Furthermore we plan to look at more limited profile lengths, based on the solution proposed by [4], and also looking at other linguistic features than character n-grams. However, focus will remain on methods that are language independent, and we plan to evaluate a recent method on multiple languages.

References

1. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (2006)
2. Kešelj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the Pacific Association for Computational Linguistics (2003)
3. Project Gutenberg Literary Archive Foundation: Free ebooks - Project Gutenberg, http://www.gutenberg.org/
4. Ruseti, S., Rebedea, T.: Authorship Identification Using a Reduced Set of Linguistic Features. Notebook for PAN at CLEF (2012)
5. Stamatatos, E.: Author Identification Using Imbalanced and Limited Training Texts. In: Proc. of the 18th International Conference on Database and Expert Systems Applications (2007)
6. Stamatatos, E.: Intrinsic Plagiarism Detection Using Character n-gram Profiles (2009)