Style-based distance features for author verification
Notebook for PAN at CLEF 2013

Erwan Moreau[1] and Carl Vogel[2]

[1] CNGL and Computational Linguistics Group
moreaue@cs.tcd.ie
[2] Computational Linguistics Group
vogel@cs.tcd.ie
Centre for Computing and Language Studies
School of Computer Science and Statistics
Trinity College Dublin
Dublin 2, Ireland

Abstract In this paper we present the approach we took in our participation to the PAN 2013 Author Identification task. It relies on a complex process to select the features which represent the author's writing, using potentially multiple statistics and distance measures computed from the training set.

1 Introduction

In this author identification task, a training set containing 35 different problems with their corresponding answer in three languages (10 in English, 20 in Greek and 5 in Spanish) is provided. Each problem consists in a small set of "known" documents by a single person and a "questioned" document; the task is to determine whether the questioned document was written by the same person.

In such an author verification task, the difficulty is the lack of negative evidence, i.e. the fact that there can be no representative corpus of text written by "any other author". To overcome this issue, our approach is inspired by the *unmasking* technique, introduced by Koppel and Schler in [2]. More precisely, we are interested in capturing the relevant features which are *unmasked* with their method, and similarly in rejecting the spurious features. However we aim to find the features which help identifying the given author *a priori*, i.e. before applying supervised learning algorithm to them. Our strategy is the following:

1. Compute a set of features based on different n-grams patterns (e.g. character trigrams, Part-Of-Speech (POS) bigrams, etc.). Each feature represents the distance between the unknown document and the author's style for this n-grams pattern.
2. For every language, feed a classification algorithm with this set of features for all the instances. Each *task* in the training set, that is, each set of documents known to have been written by a given author together with the target unknown document, corresponds to an instance.

It is worth noticing that the supervised learning stage is intended to be applied to a set of pre-selected features, which are supposed to capture individually the probability (in a broad sense) that the unknown document was written by the given author. The goal of the training stage is thus only to measure the individual contributions of the features and combine them in an optimal way. We choose this strategy because:

– The good results of the *unmasking* approach show that the key to solving this task lies in distinguishing between the n-grams which actually characterize the author and the ones which are rather specific to a particular document.

– The training set provided contains only a small set of cases (10 for English, 20 for Greek and 5 for Spanish). Thus we want to avoid using many features in the supervised training stage in order to avoid model overfitting.

We present how the features (distance values) were computed in §2. Then in §3 we explain how different models were trained and how the final ones were selected. Finally we analyze the results in §4.

2 Features

2.1 author-specific n-grams

We consider a fixed set of 14 n-grams patterns which contains tokens unigrams and bigrams, characters 4-grams, POS[3] unigrams to trigrams, plus several combinations of tokens and POS, some of which including skip-grams. For each pattern, we aim to select the set of n-grams which is the most likely to characterize the author's style.

We have observed that the more frequent a particular n-gram is, the most likely it is to follow a normal-shaped distribution accross documents by the same author.[4] This is why we use various statistics applied to the (relative) frequency of each n-gram, such as the mean, standard deviation, median and other quantiles, but also for instance the difference between the minimum and maximum or between first and third quantile. Such values are expected to provide a range against which an observed value can be compared in order to quantify how close the use if this n-gram in the unknown document is w.r.t the author's style. For each n-grams pattern, the selection of the potentially representative subset of n-grams is done by:

1. Filtering the n-grams based on one of the statistics above. A typical fitering step would be to select the n-grams for which the minimum frequency by document is higher than some threshold $t > 0$, but a few other possibilities have been tested.
2. Selecting the n-grams corresponding to the N highest or lowest values for one the statistics above. For instance the n-grams which have the smallest range between the first and third quartile are expected to characterize the author's style in the sense that the author's use of these n-grams is rather stable accross documents, while in the same time excluding possible outliers in the distribution.

We have also tried to use negative evidence by taking into account how the distribution of a selected n-gram for the given author differ from its distribution in documents written by other authors. This was done by comparing it to the each of the other authors cases in the trainining set, computing a value which represent how different the two distributions are (several methods were tested), and using the average value as criterion

[3] Part-Of-Speech tagging was done using TreeTagger (http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger) for English and Spanish, and the AUEB tagger for Greek (http://nlp.cs.aueb.gr/software.html).

[4] It is worth noticing that here we consider the frequency of a given n-gram accross different documents, independently from the other n-grams. This observation must also be taken with care because normality tests are not very reliable with small samples (here at most 10 distinct documents by the same author). Nevertheless the clear relation between frequency and normality accross documents shows that the assumption holds in general at least for frequent n-grams.

for selecting the n-gram or not.[5] This approach gave good resuts but did not bring an improvement over using only data from the author. This is why we ended not using it, since it is more complex and significantly more costly in computation time.

2.2 Comparing a document to an author profile

With the above method we can select a set of n-grams whose frequency distributions are supposed to represent the author's style. The value which will be used as feature in the supervised training stage is a distance between the questionned document and the author's style, as represented by these n-grams. Other n-grams in the unknown document are ignored, but their cumulated global frequency is indirectly taken into account in the frequencies of the selected n-grams.

Various classical distance measures have been used, like Euclidean, Cosine, χ^2, but also some ad-hoc measures which assume that the reference distribution is normal: for instance the probability of the frequency in the unknown document to belong to this distribution according to the Cumulative Distribution Function, or the simple difference between this frequency and the mean, as well as other variants involving the ranges between quantiles. Additionally it was possible to compute the final value for these ad-hoc measures according to different means: arithmetic, geometric or harmonic.[6]

3 Models training

In the following we call *distance configuration* a unique set of parameters which describe a selection and a distance method, such that applying the different steps described by these parameters to a task (set of known documents and questioned document) gives only one final value (which can be used as the value of the feature for this task/instance). Such parameters include for example the threshold and the statistic to which it is applied for a filtering step, or a distance identifier and possibly its corresponding parameters for a distance method. In order to select the best selection and comparison methods, a wide set of possible configurations have been tested.

A small set of 17 "best distance configurations" has been obtained through an incremental semi-manual evaluation based on the individual performance of the configurations: since each configuration gives a distance value for each task, it can be evaluated simply by computing the distances for all task (by language) in the training set, and then computing an optimal threshold to separate the Yes/No answers.[7] A manual analysis was carried out to assess the contribution of the various parameters, which lead to the selection of the final best distance configurations. Finally the supervised learning stage was applied to a few thousands of randomly chosen *global configurations* specified by:

- a random subset of features/n-grams patterns;
- for each pattern in the subset, a random distance configuration selected randomly from the set of 17 best distance configurations;

[5] Thus the fact that some authors appear several times in the dataset does not matter, since the impact on the average value is limited and is used only to compare n-grams from the same author (hence even if there is a bias, it is the same for all comparable values).

[6] It turned out that the arithmetic mean was less often the optimal choice than the two others.

[7] This is similar to using the correlation between the distance and the binary answer in order to compare configurations against each other, except that the result here is a maximum accuracy (more informative).

– A classification algorithm with its parameters, selected randomly from a set of 20 possible cases. The possible algorithms are SVM [1], logistic regression [3], decision trees [4] and Naive Bayes, with variants depending on their parameters.

Each random global configuration is used to produce the corresponding features and is evaluated on the training set using cross-validation Finally for each language the best performing global configuration and its corresponding model has been used in the submitted version of the software.

4 Results and discussion

19 teams participated in the competition on author identification. The following table summarizes how our system performed:

Language	F1-score	Best F1-score	Rank
English	0.767	0.800	3rd (tie with 1)
Greek	0.433	0.833	16th
Spanish	0.600	0.840	10th (tie with 4)
Global	0.600	0.753	11th (tie with 1)

Our approach performed noticeably well on English, but very bad on Greek, and in the average for Spanish. At the time of writing we cannot analyze the disappointing results on Greek, which are rather surprising since this was the biggest part of the training set (thus overfitting was less likely than with the other languages). This might be due to some technical or design problem with the POS tagger, which is the main difference compared to the two other languages.

More generally the approach is probably sensitive to overfitting, especially when trained on a small number of instances as it is the case with the training set. There are also other potential flaws which might cause an accuracy drop:

– The semi-manual features selection process might not be optimal: it relies on predefined possible parameters, and it is evaluated only on the basis of individual distance configurations, thus possibly discarding relevant combinations of features.
– The selection of the best configuration (including the set n-grams selected for an author) is a supervised process. Even if it is more indirect that the last stage of supervised learning, there might be some overlap in the information used in both stages, which could be a cause of overfitting, despite the use of cross-validation.

We intend to study these issues as future work.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the Centre for Next Generation Localisation (www.cngl.ie) funding at Trinity College, University of Dublin.

References

1. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to platt's SMO algorithm for SVM classifier design. Neural Comput. 13(3), 637–649 (Mar 2001)
2. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research 8, 1261–1276 (2007)
3. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. Mach. Learn. 59(1-2), 161–205 (May 2005), http://dx.doi.org/10.1007/s10994-005-0466-3
4. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)