Style-based distance features for author profiling

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 Profiling task. It is an adaptation of our system submitted for author identification, assuming that a profile category (authors belonging to the same gender and age group categories) can be analyzed in the same way as an author's style.

1 Introduction

In this author profiling task, we are provided with a training set of 236,000 authors for English and 75,900 authors in Spanish. We are given the gender (two categories) and the age group (three categories) of the author. The task consists in predicting the gender and age group of any new document.

Our participation in this task is an adaptation of the one we did for the author identification task[3]. Thus our participation was intended as a test to see if the two tasks can be tackled in a similar way, rather than a proper specific approach to solve this task. Nevetheless we were aware that there are major differences between the two tasks:

- The dataset consists in very noisy data, including HTML tags and various other problems (for example some documents contain only noise); some documents seem to consist in commercials or spam, which is likely not to reflect the writing style of the supposed author;
- The dataset is very big;
- More importantly, this is classical classification task with 6 possible labels, all of which being represented in the data, as opposed to the author identification task where no negative evidence can be used;
- Our approach relies on the assumption that a category (gender and age) can be treated in the same way as a single author. In particular it relies on the fact the distribution of a given n-gram among different documents in the same category is regular (in general); this means that our approach can not (at least it is not intended to) take into account several subgroups in a given category.

[3] See our paper in the same volume.

Similarly to what we did in the author identification task, we aim to compute fine-grained features which correspond to distances between the unknown document and a reference category according to a particular n-gram pattern (e.g. POS trigrams). Only one such feature at most can be used for the same pair category/pattern in the final set of features which is provided to the supervised learning algorithm.

In §2 we detail how the potential features are computed; then in §3 we explain how we had to settle for non optimal final models due to some major issue in our system; finally we present and discuss our results in §4.

2 Features

We consider a fixed set of 15 n-grams patterns which contains tokens unigrams, bigrams and trigrams, characters unigrams, trigrams and 5-grams, POS[4] unigrams to trigrams, and 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 category.

2.1 Categories

We call categories the six target profiles (two classes for gender $\times$ three classes for age group). Additionally we also consider the "superset categories" which consist in all the writers in the same gender category and all the writers in the same age category.

A very basic pre-processing is applied to the data, which consists mainly in removing any meta-data (HTML tags etc.). Each category is represented by the documents of all authors belonging to this category. However we had to restrict the data in every category to a randomly selected subset of at most 5,000 authors for efficiency reasons. Moreover, for the same reason, we had to keep only at most 12,000 distinct n-grams by discarding the least frequent ones in the category if needed.

2.2 Selecting category-specific n-grams

We had observed in the author identification task 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. We assume, maybe wrongly, that this also holds accross documents in the same category (at least for some n-grams). 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 category.

For each n-grams pattern, the selection of the potentially representative subset of n-grams is done by:

[4] 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).

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 category in the sense that for the authors in this category, the use of these n-grams is rather stable accross documents, while in the same time excluding possible outliers in the distribution.

We have also used the other categories into account by measuring how the distribution of a selected n-gram for the given category differ from its distribution in documents in the other categories. This was done by comparing the distributions using simple measures like average overlap/difference and the Bhattacharrya distance [1] (and a few variants).

2.3 Comparing a document to a category

With the above method we can select a set of n-grams whose frequency distributions are supposed to represent the category. The value which will be used as feature in the supervised training stage is a distance between the questionned document and the category, 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 (which are lower if there are many/frequent non-selected n-grams in the document for example).

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.

3 Training: what was planned and what was actually done

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 (a category as reference data and an input document) gives only one final value (which can be used as the value of the feature for this 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.

The supervised learning stage consists in specifying a *global configuration* with the following parameters:

- a random subset of pairs category/n-grams patterns;
- for each pattern in the subset, a random distance configuration selected randomly;

– A classification algorithm with its parameters, selected randomly from a set of 20 possible cases. The possible algorithms are SVM [3], logistic regression [4], decision trees [5] and Naive Bayes [2], with variants depending on their parameters.

Initially we intended to use an ad-hoc algorithm in order to select the best global configuration. This algorithm follows the principle of genetic algorithms, that is, gets incrementally closer to the optimal parameters by selecting a population at first randomly and then favoring the parameters which gave the best results in the next generation. But the algorithm failed to converge to an optimal solution. Due to the time constraints, we had to settle for the best configuration that the algorithm had found then, although it was unlikely to be optimal.

4 Results and discussion

21 teams participated in the author profiling task. Our system obtained an accuracy of 0.2395 in English (rank 19, best system: 0.3894) and 0.2539 in Spanish (rank 14, best system: 0.4208).

As explained in §3, these bad results are not surprising since the system that we submitted runs a configuration which is only the best case among random configurations. At the time of writing we have not fully investigated this yet. We think that this is probably a bug that we did not find, but do not exclude the possibility of a design flaw.

Additionally there are several other potential problems with our approach:

– The initial assumption to consider a category in the same way as a single author: if there is no such consistency among the authors who belong to the same category, our features are unlikely to work very well;
– Because of the inefficiency of our prototype, we had to ignore large parts of the training data in building the categories, which can also explain a loss in precision.
– The noisiness of the data and the unsophisticated cleaning step.

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. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of Cal. Math. Soc. 35(1), 99–109 (1943)
2. John, G., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338–345. Morgan Kaufmann (1995)
3. 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)
4. 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
5. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)