Ensembles of Proximity-Based One-Class Classifiers for Author Verification

Notebook for PAN at CLEF 2014

Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios

Faculty of Computer Science, Dalhousie University
{jankowsk, vlado, eem}@cs.dal.ca

Abstract We use ensembles of proximity based one-class classifiers for authorship verification task. The one-class classifiers compare, for each document of the known authorship, the dissimilarity between this document and the most dissimilar other document of this authorship to the dissimilarity between this document and the questioned document. As the dissimilarity measure between documents we use Common N-Gram dissimilarity based on character or word n-grams.

1 Introduction

We describe our submission to the task of Author Identification of the PAN 2014 competition [5]. This task presents participants with author verification problems, formulated as follows: “Given a small set (no more than 5, possibly as few as one) 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 who wrote the known document set.”

The required output in the competition task is a real number in the range from 0 to 1, encoding the probability of the positive answer to this question. A probability score that is less than 0.5 is interpreted as a negative answer; a probability score that is greater than 0.5 is interpreted as a positive answer; the score of 0.5 is interpreted as the "I don’t know" answer.

The submissions are evaluated using the measure of area under the ROC curve (AUC) based on the probability scores, and the c@1 measure [6]. c@1 is equivalent to accuracy when the "I don’t know" answer is not used. For a given number of correct answers, the higher number of incorrect answers is replaced by "I don’t know", the higher is c@1. The final evaluation score in the competition is the product of AUC and c@1.

The Author Identification at PAN 2014 is similar to the Author Identification task at PAN 2013, described in [2].

2 Methodology

We use an ensemble of our proximity-based one-class classifiers. The method is described in detail in [1]. For the purpose of self-containment we describe our algorithm below.

Let $A = \{d_1, ..., d_k\}$, $k \geq 2$, be a set of "known" documents written by a given author. Let u be the questioned document which authorship we are to verify.

Our algorithm calculates for each known document d_i the maximum dissimilarity between this document and all other known documents $D^{max}(d_i, A)$ as well as the dissimilarity between this document and the questioned document $D(d_i, u)$, and finally the dissimilarity ratio $r(d_i, u, A) = \frac{D(d_i,u)}{D^{max}(d_i,A)}$. We apply a threshold θ on the value of $M(u, A)$ that is the average of the $r(d_i, u, A)$ over all known documents $d_i, i = 1, ..., k$. We classify u as written by the same author as known documents iff $M(u, A) <= \theta$. Specifically, we linearly scale the average dissimilarity ratio $M(u, A)$ using the threshold θ, so that the value of M equal to θ corresponds to the score 0.5, values greater than θ correspond to the scores between 0 and 0.5, and values less than θ correspond to the scores between 0.5 and 1 (a cutoff is applied, i.e. all values of $M(u, A) < \theta - \textit{cutoff}$ are mapped to the score 1, and all values of $M(u, A) > \theta + \textit{cutoff}$ are mapped to the score 0).

For the dissimilarity measure between documents we use the Common N-Gram (CNG) dissimilarity, proposed by Kešelj et al. [4]. For each document a sequence of the most common n-grams (of characters or words) coupled with their frequencies (normalized by the length of the document) is extracted; such a sequence is called a *profile* of the document. The dissimilarity between two documents of the profiles P_1 and P_2 is defined as follows:

$$D(P_1, P_2) = \sum_{x \in (P_1 \cup P_2)} \left(\frac{f_{P_1}(x) - f_{P_2}(x)}{\frac{f_{P_1}(x) + f_{P_2}(x)}{2}} \right)^2 \qquad (1)$$

where x is an n-gram from the union of two profiles, and $f_{P_i}(x)$ is the normalized frequency of the n-gram x in the the profile P_i, $i = 1, 2$ ($f_{P_i}(x) = 0$ whenever x does not appear in the profile P_i).

If there is only one known document, we cut it in half to obtain two known documents. We also truncate all documents in a given problem to the length of the shortest one. We also make sure that each profile for a given problem has exactly the same length in cases when the number of distinct n-grams in any of the documents within given problem is less than the requested length of the profiles.

Ensembles comprise of such classifiers that differ between themselves in at least one of the following parameters: type of the tokens in n-grams (characters or words), the length of n-grams, the length of profiles. We used ensembles with weighted voting [1] in the competition submission. The output probability score of an ensemble is an arithmetic average of the scores of the single classifiers.

3 Selection of classifiers using training data

We select classifiers for the ensembles separately for each corpus, based on their performance on the training datasets. We investigate performance of classifiers, varying their parameters. The tokens were utf8-encoded characters or turned to uppercase words. For classifiers based on characters the length of n-grams varied from 3 to 10. For classifiers based on word n-grams the length of n-grams varied from 1 to 6. The length of profiles was in $\{200, 500, 1000, 1500, 2000, 2500, 3000\}$ for both kinds of tokens. This space of parameters results in 98 single classifiers: 56 character-based ones and 42 word-based ones. Package Text::Ngrams [3] has been used in the software. For scaling of M values to the probability scores the cutoff was set to 0.2.

We select for each training corpora separately 31 classifiers that yield the best AUC. Subsequently for each of those classifiers the optimal threshold is found (i.e., the threshold for which the maximum accuracy is achieved). In an ensemble for a givne corpus, the threshold for all classifiers is set to one value: the average of the optimal thresholds on the training data for the selected single classifiers.

The ranges of AUC and of maximum accuracy (accuracy at the optimum threshold) for the sets of 31 classifiers are presented in Table 1.

	range of results of 31 classifiers with the highest AUC		parameters of the classifier with the highest AUC		
training corpus	AUC	maximum accuracy	token	n-gram length	profile length
Dutch essays	$0.82 - 0.86$	$0.77 - 0.83$	character	5	500
Dutch reviews	$0.54 - 0.56$	$0.55 - 0.59$	character	7	200
English essays	$0.52 - 0.55$	$0.52 - 0.58$	word	4	3000
English novels	$0.64 - 0.74$	$0.62 - 0.71$	word	1	500
Greek articles	$0.68 - 0.79$	$0.66 - 0.77$	word	1	500
Spanish articles	$0.82 - 0.85$	$0.76 - 0.80$	word	1	500

Table 1. Results of experiments on the training corpora in the PAN 2014 competition task Author Identification.

We observe that our method performs best for the training corpus for Dutch essays and Spanish articles. It performs worse on the Greek articles set. For the sets of English novels and Dutch reviews the performance is low. Most likely the reason behind that lies in the fact that in these two sets all but one problem have exactly one known document. We observed that such problems are especially challenging for our method. This is most likely because the two halves of a single known document, that we compare the questioned document with, are much more similar to each other than two different documents written by the same person. The reasons behind the low results on the English essays set are not clear to us and require further investigation.

4 Competition results

The results of our submission on the PAN 2014 evaluation set for the Author Identification tasks are presented in Table 2.

	AUC	c@1	final score
Dutch essays	0.86892	0.84201	0.73165
Dutch reviews	0.6376	0.56	0.35706
English essays	0.5179	0.54837	0.284
English novels	0.49125	0.45727	0.22464
Greek articles	0.7308	0.68	0.49694
Spanish articles	0.8026	0.73	0.5859

Table 2. The results of our submission in the PAN 2014 competition task Author Identification (as announced on June 11, 2014). The final score is the product of the values of AUC and c@1.

References

1. Jankowska, M., Milios, E., Kešelj, V.: Author Verification Using Common N-Gram Profiles of Text Documents. In: Proceedings of the 25th International Conference on Computational Linguistics. COLING '14 (August 2014)
2. Juola, P., Stamatatos, E.: Overview of the Author Identification Task at PAN 2013. In: Forner, P., Navigli, R., Tufis, D. (eds.) Working Notes Papers of the CLEF 2013 Evaluation Labs (September 2013)
3. Kešelj, V.: Perl Package Text::Ngrams (2013), http://www.cs.dal.ca/ vlado/srcperl/Ngrams
4. 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'03. pp. 255–264. Dalhousie University, Halifax, Nova Scotia, Canada (August 2003)
5. PAN: Pan competition, author identification (2014), http://www.uni-weimar.de/medien/webis/research/events/pan-14/pan14-web/author-identification.html
6. Peñas, A., Rodrigo, A.: A simple measure to assess non-response. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 1415–1424. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (June 2011)