A Single Author Style Representation for the Author Verification Task

Notebook for PAN at CLEF 2014

Cristhian Mayor[1,3], Josue Gutierrez[3], Angel Toledo[3], Rodrigo Martinez[3], Paola Ledesma[2,3], Gibran Fuentes[3], and Ivan Meza[3]

[1]Institut National des Sciences Appliquees de Lyon (INSA Lyon)
http://www.insa-lyon.fr/
[2]Escuela Nacional de Antropologia e Historia (ENAH)
http://www.enah.edu.mx
[3]Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas (IIMAS)
Universidad Nacional Autonoma de Mexico (UNAM)

Abstract This paper presents our experience implementing three approaches for the 'PAN 2014 Author Identification' [3,1] task using the same representation for the author's style. Two of our approaches extend previous successful approaches: naive Bayes [4] and imposter [8] methods. The third approach is based on original research on sparse representation for text documents. We present results with the official development and test corpora.

1 Introduction

Author verification has multiple applications on several areas including information retrieval and computational linguistics, and has an impact in fields such as law and journalism [2,5,9]. In this edition of the *PAN 2014 Author Identification*, the task was formally defined as follows[1]:

> *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.*

This year the documents were in four languages and four genre with the following combinations: essays and reviews in Dutch, essays and novels in English and articles in Greek and Spanish.

In this work we present the implementation of three approaches to perform the authorship verification task based on the same document representation. In particular, we continue using a vector space representation for documents as presented in our approach last edition [6], but we agglomerate them to obtain a single author's style representation. This representation was used on naive Bayes [4] and imposter [8] commonly known methods. Additionally, we implemented a novel approach using sparse representation [10].

[1] As described in the official website of the competition http://pan.webis.de/ (2014).

2 Author's Style representation

The representation for an author's style is generated in two stages. First, we represent the documents from an author using the vector space model to represent a document[7]:

$$d = (w_1, w_2, \ldots, w_m)$$

In which w_i is a frequency or weight of a *word* of the vocabulary of size m for the word i. A common representation of a vector space model is the bag of words model, in which the words represent actual words of the document and the frequencies counts of occurrence of such words in the represented document.

Additionally to bag of words we use the following feature frequencies to represent the documents:

Bag of words Frequencies of words in the document.
Bigram Frequencies of two consecutive words.
Trigram Frequencies of three consecutive words.
Prefix Frequencies of prefixes of words.
Suffix Frequencies of suffixes words.
Prefix bigram Frequencies of two consecutive prefixes of words.
Suffix bigram Frequencies of two consecutive suffixes words.
Stop words Frequencies of stop words.
Stop words bigram Frequencies of two consecutive stop words.
Punctuation Frequencies of punctuations.
Words per sentence Frequencies of words per sentence.

In the second stage an author is represented by the sum of the vectors representing the documents written by her or him. This accumulative vector is normalized by the number of the documents by the author and it represents the style of the author. This representation is not novel however many approaches on author verification optimize the representation on the domain, in our experiments we keep the same representation independently of the domain. Some of the implemented approaches require an instance of the document, for this it is possible to sample a document from the author's style representation.

3 Approaches

The approaches implemented in the task are described next.

3.1 Imposters

This method consists on iteratively compare the vector distance between the author's document to the questioned document versus the distances between several imposter documents to the questioned document. With these distances a score is built up based on how many times the author and questioned documents are closer than the imposter and questioned documents. For this approach we follow the description of the method by *Seidman* (2013) [8]. We modify the method to work on more than one set of features and instead to use imposters from the web we used the training corpus as source of imposters. Additionally, we extended the approach to produce a probability as output based on repetition of the algorithm.

3.2 Naive Bayes

This method consists on sampling from the author and the imposter style representation two documents instances for each. A probability score is then calculated using the common term between the questioned document and the authors documents. On the other hand, an alternative score is calculated between the questioned document and the impostor documents. These scores are derived using Bayes[2]. The purpose of the score is to capture the probability that the document was created by the same author, if the score for the author is higher than the imposter we consider as evidence of authorship. We iterate n times over this method to calculate the probability of authorship.

3.3 Sparse

This methodology has been successful in the face recognition task [10]. We adapted this methodology to work in the authorship verification task. The method consists on identifying the components of the questioned document from samples (a dictionary) of documents from authors. In theory, the biggest contribution of components had to be by elements of a single author. In order to identify the components the method proposes the following l^1-minimization:

$$x_1 = argmin||x||_1 \text{ subject to } Ax = y$$

On which, y is the questioned document, A the matrix of n samples from different m candidate authors (imposters), and x the variable to minimize which represent the contribution from each candidate. From the resulting variable x_0 we can quantify the residuals given by Ax versus y and decide which author contributes with more components. We adapt this method to produce a probability as result by iterating k times over the full method.

4 Results

Table 4 presents the overall development results reached with each one of the approaches.

Approach	AUC	C@1	Score
imposter	62%	56%	35%
n-gram	64%	57%	36%
Sparse	72%	68%	48%

Table 1. Overall results for all approaches.

The results for language and genre for our best system (i.e., sparse)are presented in Table 4.

[2] This method is inspired in the following document: http://cs229.stanford.edu/proj2009/Leahy.pdf.

Approach	AUC	C@1	Score
Dutch reviews	93%	88%	82%
Dutch essays	57%	52%	30%
English essays	57%	56%	32%
English novels	66%	61%	41%
Greeks articles	82%	75%	62%
Spanish news	75%	71%	54%
Overall	70%	65%	50%

Table 2. Detailed final scores for language and genre for the *sparse* approach on testing corpus.

5 Discussion

In the preparation of authorship verification system we implemented three approaches: *imposter*, n-*gram* and *sparse* methodologies. During development we tested all of them on the same representation for the author's style for the four languages and three genre of the task. During development and testing the best results were reach using the *sparse* methodology which is interesting to us since it was the first time such method was applied to the task of authorship verification.

References

1. Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: irecent trends in digital text forensics and its evaluation. in pamela forner. In: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative (CLEF 13) (2013)
2. Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1(3), 233–334 (Dec 2006)
3. Juola, P., Stamatatos, E.: Overview of the author identification task at pan 2013. In: CLEF 2013 Evaluation Labs and Workshop - Online Working Notes (2013)
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. vol. 3, pp. 255–264 (2003)
5. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology 60(1), 9–26 (2009)
6. Ledesma, P., Fuentes, G., Jasso, G., Toledo, A., Meza, I.: Distance learning for author verification. In: Proceedings of the conference pacific association for computational linguistics, PACLING. vol. 3, pp. 255–264 (2003)
7. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (Nov 1975)
8. Seidman, S.: Authorship verification using the impostors method. In: CLEF 2013 Evaluation Labs and Workshop - Online Working Notes (2013)
9. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009)
10. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(2), 210–227 (2009)