Authorship verification, combining linguistic features and different similarity functions

Notebook for PAN at CLEF 2015

Daniel Castro[1], Yaritza Adame[1], María Pelaez[1], Rafael Muñoz[2]

[1]Desarrollo de Aplicaciones, Tecnología y Sistemas
DATYS, Cuba
{daniel.castro, yaritza.adame, maria.pelaez}@datys.cu
[2]Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante, España
rafael@dlsi.ua.es

Abstract. Authorship analysis is an important task for different text applications, for example in the field of digital forensic text analysis. Hence, we propose an authorship analysis method that compares the average similarity of a text of unknown authorship with all the texts of an author. Using this idea, a text that was not written by an author, would not exceed the average of similarity with known texts and a text of unknown authorship would be considered as written by the author, only if it exceeds the average of similarity obtained between texts written by him and if it got the major value comparing the average similarity with the rest of the authors. For each linguistic feature we obtain a vote by majority using different functions and for the final decision we divide the number of votes for each feature that consider as written by the author the unknown text by the total of features analyzed. The results obtained for each language in the PAN 2015 authorship verification competition are exposed in the overview of the task.

Keywords: Authorship detection, Author identification, similarity measures, linguistic features.

1 Authorship analysis

To determine the true author of a document has been a task of social interest from the moment it was possible to attribute the authorship of words. Questions about the authorship of a document may be of interest not only to specialists in the field (forensics specialist, linguistics researchers, etc.), but also in a much more convenient sense for politicians, journalists, lawyers. Recently, with the development of statistical techniques and because of the wide availability of accessible data from computers, the authorship analysis automatically has become a very practical option. [5]
There are many practical examples where the authorship analysis becomes the key to solve them. Suppose a malicious mail is sent using an email account belonging to someone else, who subsequently is accused of this fact. Who is the author of the mail?

It may happen that a person dies and there is a note that makes it seem that the person committed suicide. Was it really a suicide note or was it used to cover up a murder? [5] There may be a document, for example a digital newspaper that could have been altered so it couldn't be used as evidence in a trial. Was this newspaper altered or not?
The authorship analysis task affronts the problem of determining the author of an anonymous document or one whose author is in doubt. For this it is necessary to try to infer linguistic characteristics (features) of the author through documents written by him, features that will allow us to create a model of the writing style of this author and measure how similar may be any unknown document to documents written by that author.
One of the principal evaluation labs for the dissemination, experimentation and collaboration in the development of methods for the authorship analysis is found in the PAN evaluation forum[1]. It's important to notice, that most of the papers presented in different editions of this evaluation forum [6, 10] used Natural Language Processing tools, in order to obtain the linguistic features which identify an author and differentiate it from the rest.
In PAN, 2013 and 2014 editions, specifically it was tested the task of authorship verification, where authors samples are formed by documents of a known author and an unknown document to check whether it was written by that author. No restrictions are imposed on the use of samples of others for support in finding a decision, or just use the samples of a single author, the latter idea would be challenging and difficult.
The basic properties of the papers presented in the PAN 2014 [10] evaluation are:

- By the use of known documents samples of authors: intrinsic (only the documents of the author in analysis) or extrinsic (using samples of others authors).
- Type of machine learning algorithms or approximation used: lazy or hard-working approaches (more training computational costs).
- Type of linguistic features used: low-level features (characters, phonetic and lexical) and/or syntactic.

1.1 Linguistic features

The linguistic features are the core of the authorship analysis task (regardless of the subtask or approach used in the analysis, such as author verification, author detection, plagiarism detection, etc.), they can be used to coded documents with any mathematical model, traditionally being the vector space model the approximation most used. The purpose lies in trying to identify a writing style of each author to distinguish it from the rest [5].
There are several number of features that have been taken into account in the authorship analysis task, in the majority is used a distribution of features grouped by linguistic layers (we call them also features obtained from the content writing) [1, 4, 7, 8].
Five linguistic feature layers are identified in [11]: phonetic, character, lexical, syntactic and semantic layer:

[1]http://pan.webis.de

1. Phonetic layer: This layer includes features based on phonemes and can be extracted from the documents through dictionaries. Example: the International Phonetic Alphabet (IPA).
2. Character layer: This layer includes character-based features as prefixes, suffixes or n-grams of letters.
3. Lexical layer: This layer includes features based on terms such as auxiliary words.
4. Syntactic layer: This layer includes syntax based features such as sentences components.
5. Semantic layer: This layer includes semantic-based features as homonyms or synonyms.

Based on this structure feature layers, in our present work we use features of the 2, 3 and 4 layers, which we illustrate in more detail in next sections.
In Section 2 we present the characteristics of our method and in section 3 the experimental results obtained in the PAN 2015 authorship verification task. Finally conclusions and future work.

2 Combining linguistic features and different similarity functions

There are various aspects that need to be analyzed in order to implement a method that allows us to assess whether a text of unknown or disputed authorship, was written by an author from which we have written text samples. It should be considered whether samples of the author belong to the same genre, theme, were written with a considerable time difference, are written in the same language or have sections written in other languages, or if the samples have been revised and corrected by someone else.
Our method is based on the analysis of the average similarity (AS_{Unk}) of an unknown authorship text with the closeness to each of the samples of an author, comparing it to the Average Group Similarity (AGS) between samples of an author.
We performed experiments with a total of 10 types of linguistic features (we will illustrate the features in the following section) and used three similarity functions.
We identified three key steps in our method, these are:

1. Representation of all documents by one feature type.
2. Average similarity between the document samples of an author (AGS).
3. Average similarity between the document of unknown authorship and the known samples of each author in a set (AS_{Unk}), in which we know who is the author that is been analyzed and the rest are used as impostors [9].
4. For each linguistic feature analyzed, we obtain a vote by majority combining the use of different similarity functions, in which 1 represents that the document was written by the author in analysis and 0 the opposite.
5. We obtain as a final decision a value in the [1, 0] interval, dividing all the votes with 1 for the features by the total number of features used, in this case the number of features used is 10.

2.1 Linguistic features used to represent the documents

We use the vector representation to store the values of the linguistic features extracted form one document, so each sample (document) with known or unknown author is represented by 10 vectors corresponding to each of the types of features with which experiments were performed.
The features evaluated and calculated are grouped in three layers: character, word and syntactic (lemma and Part of Speech)

1. Character
   (a) Tri-grams of characters (F1)
   (b) Quad-grams of characters (F2)
   (c) Word prefixes of size 2 (F3)
   (d) Word suffixes of size 2 (F4)
2. Words
   (a) Uni-grams of words (F5)
   (b) Tri-grams of words (F6)
3. Lemma and Part of Speech
   (a) Uni-grams of lemmas (F7)
   (b) Uni-grams of Part of Speech (F8)
   (c) Tri-grams of lemmas (F9)
   (d) Tri-grams of Part of Speech (F10)

The features of the third layer of analysis are obtained using tools of Natural Language Processing implemented in the Xinetica[2] platform.

2.2 Average similarity

We show in the next picture (Figure 1) the process to calculate the average similarity of the documents of the known author and the average similarity of these samples with the unknown text. Initially we have several samples of documents (Doc) by an author and a document of unknown authorship (Unk).
The first task is to represent each of these documents in a vector space model, analyzing one type of feature. Subsequently, for the document samples of the known author, we analyze the average similarity of each document with the rest, using the following formula:

$$AS_j = \frac{\sum_{O_j \in K_j} Sim(O, O_j)}{|K_j| - 1} \quad (1)$$

Where "O" would be a document of the author and "O_j" the rest of the documents of the same author, K_j represents the author and $|K_j|$ the number of documents of the author. By $Sim(O, O_j)$, is represented the similarity between two documents.
Therefore, for each known document of the author their average similarity with the other is calculated and finally, the average similarity of all samples is calculated or what we call the Average Group Similarity (AGS):

[2] http://www.cerpamid.co.cu/xinetica/index.htm

$$AGS = \frac{\sum_{o_j \in K_j} AS_j}{|K_j|} \quad (2)$$

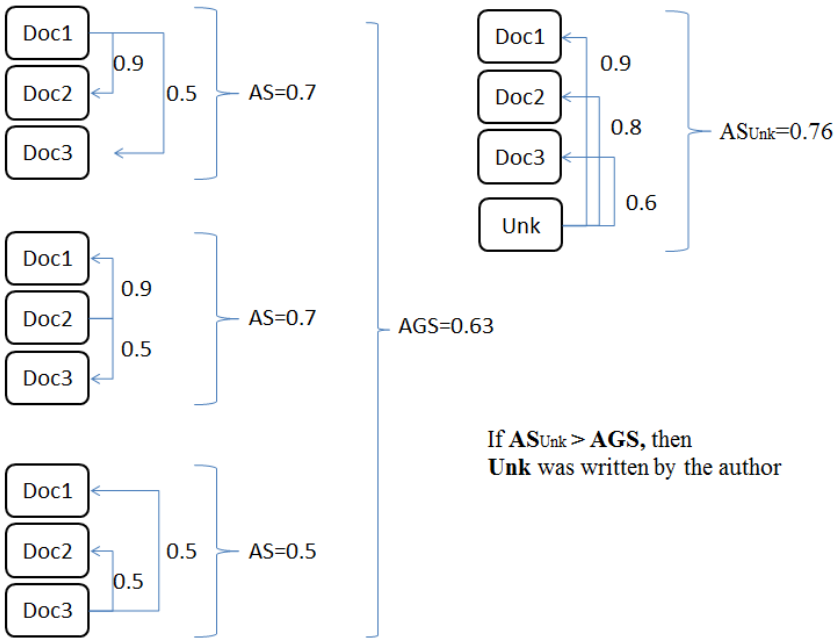

Fig. 1. Average Group Similarity (AGS) analysis of an author documents samples and Average Similarity (AS_{Unk}) of an unknown authorship document

Given a document of unknown authorship, initially it must be represented by the type of feature in which samples of the known author are represented in order to be compared. Then the AS_{Unk} is calculated using the known samples. The decision is made by comparing the AGS with calculated AS_{Unk} of the unknown document. If $AS_{Unk} < AGS$, then the unknown sample is not considered written by this author. To determine if the response is positive (that is, that the document of unknown author was written by the author of the given samples), then $AS_{Unk} \geq AGS$ and the AS_{Unk} obtained with the author in analysis must be the highest. The set of impostors authors used are the set of authors of the test corpus.

We have implemented 3 similarity functions in order to perform experiments with each of them, these are: Cosine, Dice and MinMax [3]. For the MinMax function, we use as similarity 1-MinMax.

We focus then our study in analyzing two aspects:

1. The idea of the AGS measure as a limit to determine when an unknown document was written by an author. This could be a strict limit to determine when a text was written by an author.
2. Take a **final-decision** based on the combination of the results of pair function-feature for each linguistic feature, and all the decisions using the total number of features. See Figure 2.

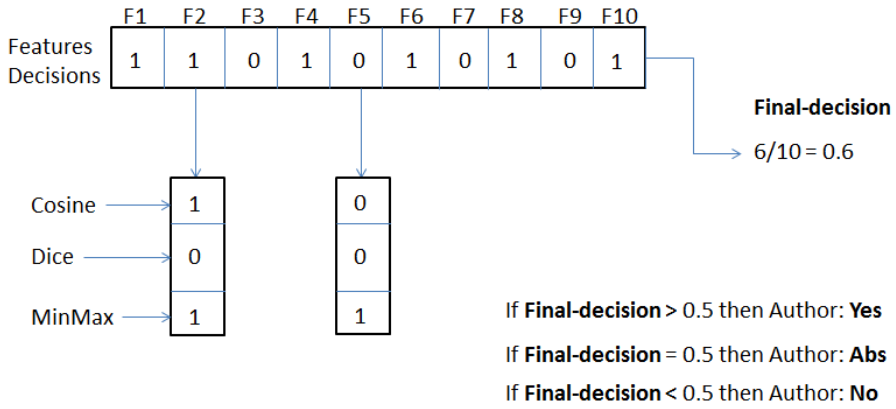

Fig. 2. Combining functions in each feature and getting de final decisions.

3 Experimental results

This section shows the results of evaluating the method presented in the evaluation lab of the PAN 2015 authorship verification task. We run the experiments for the 4 language dataset released. The evaluation measure and the composition of the dataset are described in the overview of the competition [2].

We tested our method with the 4 dataset provided, 2 of them, Spanish and Dutch consisting of a set of authors with cross-genre samples; for English and Greek the samples were cross-topic. We achieved the best results for the cross-topic dataset, and the best was for English. In the case of the Dutch and Greek dataset, we used only 6 features because we didn't have a tagger for these languages.

In the next tables, we show the results obtained by the participants and our results are those corresponding with the name "castro15".

Table 1. Results obtained by us and the best result for each language

Language	ranking/ participants	User	AUC	C1	Final Score
Dutch	1/17	moreau15	0.8253	0.7697	0.63523
	13/17	**castro15**	0.50287	0.49091	0.24686
English	1/17	bagnall15	0.8111	0.75651	0.61361
	2/17	**castro15**	0.74987	0.694	0.52041
Greek	1/15	bagnall15	0.8822	0.8505	0.75031
	10/15	**castro15**	0.621	0.63	0.39123
Spanish	1/17	bartoli15	0.9318	0.83	0.77339
	13/17	**castro15**	0.5576	0.59	0.32898

4 Conclusions and future work

We have presented the implementation of a method for authorship analysis that compares the average similarity calculated between a document of unknown authorship and documents written by an author, with the average similarity of the samples of this author.
Using this idea, a text that was not written by an author, would not exceed the average of similarity with known texts and only the text of unknown authorship would only be considered as written by the author, if it exceeds the average of similarity obtained between texts written by him and if it has the highest value taking into consideration the rest of the authors.
To prove the idea, we use 10 types of linguistic features to represent the documents and evaluate the similarity between two vector representations of documents using one of three similarity functions implemented. We obtained the best results over the cross-topic dataset for English and Greek language.
We propose as future work: consider a text as written by the author only in case that the average similarity of the unknown text is superior than the AGS; prove as a limit to determine if the unknown text is of the author if his AS_{Unk} is superior to the less AS of one of the known document sample. Evaluate overall different genre of documents if all the features or functions contribute to the task.

5 Acknowledgements

This research has been partially funded by the Spanish Ministry of Science and Innovation (TIN2012-38536-C03-03)

6 References

1. Castillo, E. Vilariño, D. Pinto, D. León, S. Cervantes, O.: Unsupervised method for the authorship identification task. Notebook for PAN at CLEF 2014. (2014)
2. Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio Lopez Lopez, Martin Potthast, and Benno Stein. : Overview of the Author Identification Task at PAN 2015. In Working Notes Papers of the CLEF 2015 Evaluation Labs, CEUR Workshop Proceedings. CLEF and CEUR-WS.org. ISSN 1613-0073. (2015)
3. Gomaa, W. and A. Fahmy.: A Survey of Text Similarity Approaches. International Journal of Computer Applications (0975 – 8887) Volume 68– No.13. (2013)
4. Halvani, O. Steinebach, M and Zimmermann, R.: Authorship Verification via k-Nearest Neighbor Estimation. Notebook for PAN at CLEF 2013. (2013)
5. Juola, P.: Authorship Attribution. Foundations and Trends in Information Retrieval Vol. 1, No. 3 (2006) 233–334. (2008)
6. Juola, P. and E. Stamatatos.: Overview of the Author Identification Task at PAN 2013. CLEF 2013. (2013)
7. Khonji, M and Iraqi, Y.: A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF). Notebook for PAN at CLEF 2014. (2014)
8. Ruseti, S and Rebedea, T.: Authorship Identification Using a Reduced Set of Linguistic Features. Notebook for PAN at CLEF 2012. (2012)

9. Seidman, S.: Authorship verification using the impostor methods. Notebook for PAN at CLEF 2013. (2013)
10. Stamatatos, E., W. Daelemans, B. Verhoeven, M. Potthast, B. Stein, P. Juola and M. Samchez-Perez.: Overview of the Author Identification Task at PAN 2014. CLEF 2014. (2014)
11. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, 60(3), pp. 538-556, 2009, Wiley. (2009)