Unsupervised method for the authorship identification task

Notebook for PAN at CLEF 2014

Esteban Castillo[1], Ofelia Cervantes[1], Darnes Vilariño[2], David Pinto[2], and Saul León[2]

[1]Universidad de las Américas Puebla
Department of Computer Science, Electronics and Mechatronics, Mexico
E-mail addresses: {esteban.castillojz, ofelia.cervantes}@udlap.mx

[2]Benemérita Universidad Autónoma de Puebla
Faculty of Computer Science, Mexico
E-mail addresses: {darnes, dpinto}@cs.buap.mx and saul.ls@live.com

Abstract This paper presents an approach for tackling the authorship identification task. The approach is based on comparing the similarity between a given unknown document against the known documents using a number of different phrase-level and lexical-syntactic features, so that an unknown document can be classified as having been written by the same author, if the different similarity measures obtained are close to a predetermined threshold for each language in the task. The method has shown competitive results, achieving the overall 6th place in the competition ranking.

Keywords: Authorship verification, features, Similarity measures, unsupervised learning, threshold

1 Introduction

Discovering the correct features in a raw text in order to unambiguously allow to attribute the authorship of a given anonymous document is a very hard problem that recently (empowered by the continuous growing of information in Internet) has become of high interest in areas like information retrieval (IR), Natural Language Processing (NLP) and computational linguistics. Taking into account the above, the most common framework for mapping candidate authors with unknown documents is the authorship attribution problem, where, given texts of uncertain authorship and sample documents from a small, finite set of candidate authors, the task consists of mapping the uncertain texts onto their true authors among the candidates. This problem is considered as an unreasonably easy task while a more demanding problem (often presented in documents on the web) is the author verification task, where, in a given set of documents by a single author and a questioned document, the problem is to determine if the questioned document was written by that particular author or not. In this sense, the importance of finding the correct features for characterizing the signature or particular writing style of a given author is fundamental for solving the authorship problem.

The results reported in this paper were obtained in the framework of the 8th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 14), in particular, in the task named **Author Identification**. For this purpose, we have attempted an approach for representing the features that will be taken into account in the process of authorship verification using an unsupervised learning method. The proposed approaches are discussed in the following sections.

The rest of this paper is structured as follows. In Section 2 it is presented the description of the features used in the task to be tackled. Section 3 shows the Proposed approach (unsupervised algorithm) used in the experiments. The experimental setting and a discussion of the obtained results are given in Section 4. Finally, the conclusions of this research work is presented in Section 5.

2 Description of the features and similarity measures used in the task

In this work we explore a combination of different types of features in a text frecuency vector [2] to represent the writing style of the authors. For all languages in the author verification task, we consider using lexical-syntactic features because they are relatively easy to quantify as well as they provide the syntactic order in which the words are used to form the ideas. On the other hand, it is also considered (in an unsupervised classification) the use of different metrics to determine the **similarity** of the feature vectors related to the documents of known origin against the documents of unknown origin. Both types of elements are described below.

2.1 Lexical-syntactic features

In this approach are considered the following features for representing the particular writing style of a given author:

- Phrase level features
  1. **Word prefixes**. A group of letters added before a word or base to alter its meaning and form a new word.
  2. **Word sufixes**. A group of letters added after a word or base to alter its meaning and form a new word.
  3. **Stopwords**. A group of words that bear no content or relevant semantics which are filtered out from the texts.
  4. **Punctuation marks**. Conventional signs and certain typographical devices as aids to the understanding and correct reading, both silently and aloud, of handwritten and printed texts.
  5. **N-grams**. An n-gram [5] is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the speech corpus.
  6. **Skip-grams**. A technique [4] where the n-grams are used to stored sequences of words, but they allow to skip tokens.

– Character level features

1. **Vowel combination**. Word consonants are removed and, thereafter, the remaining vowels are combined. Each vowel combination is considered to be a feature. Adjacent repetition of vowels are considered as only one vowel.
2. **Vowel permutation**. Word consonants are removed and, thereafter, the vowel permutation is considered to be a feature.

2.2 Similarity measures

In this approach the following metrics are considered to determine the similarity of the documents:

– **Latent semantic analysis (LSA)**. A technique[6] which analyzes relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words which are close in meaning will occur in similar pieces of text.
– **Jaccard similarity**. A metric used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and it is defined as the size of the intersection divided by the size of the union of the sample sets and is given by:

$$JS(A,B) = \frac{A \bigcap B}{A \bigcup B} \tag{1}$$

– **Euclidean distance**. A metric used for comparing the similarity between two vectors p and q, where, $p = (p_1, p_2, ..., p_n)$ and $q = (q_1, q_2, ..., q_n)$, then the distance from p to q, or from q to p is given by:

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{2}$$

– **Chebyshev Distance**. A metric defined on a vector space where the distance between two vectors p and q, is the greatest of their differences along any coordinate dimension. The distance is given by:

$$D_{Chebyshev}(q,p) := \max_i(\mid q_i - p_i \mid) \tag{3}$$

– **Cosine similarity**. A measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. Given two vectors of attributes, p and q, the cosine similarity, is represented using a dot product and magnitude and is defined as:

$$cos(\Theta) = \frac{p \cdot q}{\parallel p \parallel \parallel q \parallel} = \frac{\sum_{i=1}^{n} p_i \times q_i}{\sqrt{\sum_{i=1}^{n}(p_i)^2} \times \sqrt{\sum_{i=1}^{n}(q_i)^2}} \tag{4}$$

Data: X: Unknown documents and $Y = \{Y_1, Y_2, \ldots, Y_n\}$ known documents.
Result: Same author or different author for each Unknown document

```
/* For each unknown document in the test data set   */
for X do
    index ⟵ 0;
    vector_1, vector_2 ⟵ [ ];
    for ⟨X, Y_i⟩ ∈ Document set do
        /* Frequency vector of all the combined Lexical
           syntactic features of the test document     */
        X_v ⟵ Feature representation of X;
        Y_v ⟵ Feature representation of Y_i;
        /* Cosine Similarity                           */
        vector_1 [index] ⟵ S_Cosine(X_v, Y_v);
        /* Frequency vector of the words of the test
           document                                    */
        X_v ⟵ Feature representation of X;
        Y_v ⟵ Feature representation of Y_i;
        /* Jaccard similarity                          */
        vector_2 [index] ⟵ S_Jaccard(X_v, Y_v);
        index ⟵ index + 1;
    end
    Cosine ⟵ max(vector_1 [0], ..., vector_1 [n]);
    Jaccard ⟵ max(vector_2 [0], ..., vector_2 [n]);
    index ⟵ 0;
    vector_1, vector_2, vector_3 ⟵ [ ];
    for ⟨X, Y_i⟩ ∈ Document set do
        /* LSA word vectors of the test document       */
        X_v ⟵ Feature representation of X;
        Y_v ⟵ Feature representation of Y_i;
        /* Cosine Similarity                           */
        vector_1 [index] ⟵ S_Cosine(X_v, Y_v);
        /* Chebyshev Distance                          */
        vector_2 [index] ⟵ S_Chebyshev(X_v, Y_v);
        /* Euclidean distance                          */
        vector_3 [index] ⟵ S_Euclidean(X_v, Y_v);
        index ⟵ index + 1;
    end
    LSA1 ⟵ max(vector_1 [0], ..., vector_1 [n]);
    LSA2 ⟵ max(vector_2 [0], ..., vector_2 [n]);
    LSA3 ⟵ max(vector_3 [0], ..., vector_2 [n]);
    MaxSimilarity ⟵ max(Cosine, Jaccard, LSA1, LSA2, LSA3)
    if MaxSimilarity >= Δ then
        result ⟵ same author
    else
        result ⟵ different author
    end
end
```

Algorithm 1: Proposed approach using an unsupervised learning method

3 Proposed approach

For tackling the authorship identification task we propose a methodology (see algorithm 1) in which we used an unsupervised learning method. We evaluate different feature vectors with different similarity measures in order to identify if an unknown document belongs to an author. For each language, three types of text representations are evaluated:

- A frequency vector of all the vocabulary in the test documents.
- A frequency vector of all the combined Lexical syntactic features of the test documents:
  1. Word prefixes
  2. Word sufixes
  3. Stopwords
  4. Punctuation marks
  5. N-grams(from 1 to 5)
  6. Skip-grams (from 1 to 5)
  7. Vowel combination
  8. Vowel permutation
- A similarity vector using the LSA algorithm for each word in the test documents.

Different distance/similarity measures were tested, including the Jaccard similarity for the vocabulary feature vector, the cosine similarity fo the Frequency vector of all the combined Lexical syntactic features and Chebyshev Distance, Euclidean distance and cosine similarity for the LSA vectors. The best similarities of each document of unknown origin were obtained and a threshold (see table 1) to determine if the document belongs to an author. The threshold was obtained using a classification tree [8] on the maximum similarities in the test data set.

Table 1. Threshold used in the authorship identification task

Language	**Genre**	**Threshold** Δ
Dutch	essays	0.82
Dutch	reviews	0.93
English	reviews	0.67
English	novels	0.80
Greek	articles	0.76
Spanish	articles	0.64

4 Experimental results

The results obtained with the approach are discussed in this section. First, we describe the dataset used in the experiments and, thereafter, the results obtained.

4.1 Data sets

The description of the six text collections used in the experiments are shown in Table 2. As can be seen, the data set is made up of different languages/genres within the author verification task, where, each problem consists of some (up to five) known documents written by a single person and only one questioned document. All documents within a single problem instance are in the same language and the documents are matched for register, theme, and date of writing.

Table 2. Data set used in the authorship identification task

Language	Genre	Number of training documents	Number of test documents
Dutch	essays	169	96
Dutch	reviews	105	100
English	reviews	529	200
English	novels	100	100
Greek	articles	100	100
Spanish	articles	500	100

4.2 Results obtained in the task

In Table 3 we present the results obtained by our approach using the TIRA tool [3] with each one of the data sets considered in the competition. The results were evaluated according to the area under the ROC curve (AUC) measure [1] and the c@1 measure [7]. Three different languages (each one, with different genres) were tackled out. The best performance was obtained with the Dutch language in the essays genre, followed by the Greek language in the articles genre, and the English language in the essays genre. However, a low performance was obtained (with respect to the other teams in the competition) in the Dutch, Spanish and English languages in the reviews, articles and novels genres. We consider these results were obtained because of the use of empirical thresholds in the final classification of the unknown documents. Further analysis will investigate this issue. It is worth to notice that we obtained the sixth place from 13 teams and that our approach always performed better than the competition baselines.

Table 3. Results obtained in different languages

Language	Genre	AUC	C@1	Final score	Runtime	Ranking based on the final score
Dutch	**essays**	**0.86068**	**0.86133**	**0.74133**	**00:01:56**	5
Dutch	reviews	0.6692	0.3696	0.24734	00:01:00	11
English	essays	0.54855	0.58	0.31816	01:31:53	11
English	novels	0.6278	0.615	0.3861	26:14:11	5
Greek	articles	0.686	0.73	0.50078	00:03:14	4
Spanish	articles	0.734	0.76	0.55784	00:06:48	5

5 Conclusions

We have presented an approach that uses an unsupervised method with lexical-syntactic features. Even if the runtime is greater than the most approaches of this competition, the performance is good. It was surprising that being a Spanish native language team, we performed better in Dutch and Greek languages, but it is a good oportunity for analyzing the text into more depth for determining the reason of this issue. As we mentioned before, we have executed the same methodology across the different languages, varying basically only the thresholds applied in the final classification. However, more experiments continue to be performed to analyze whether or not these changes introduce significant variations in the data sets. Future work is planned to observe the performance of the proposed methodology using different similarity measures.

References

1. Agarwal, S., Graepel, T., Herbrich, R., Har Peled, S., Roth, D.: Generalization bounds for the area under the roc curve. Journal of Machine Learning Research 6, 393–425 (2005)
2. Christopher D. Manning, Prabhakar Raghavan, H.S.: Introduction to Information Retrieval (2008)
3. Gollub, T., Potthast, M., Beyer, A., Busse, M., Pardo, F.M.R., Rosso, P., Stamatatos, E., Stein, B.: Recent trends in digital text forensics and its evaluation - plagiarism detection, author identification, and author profiling. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF. Lecture Notes in Computer Science, vol. 8138, pp. 282–302. Springer (2013)
4. Guthrie, D., Allison, B., Liu, W., Guthrie, L., Wilks, Y.: A closer look at skip-gram modelling. In: Proceedings of the Fifth international Conference on Language Resources and Evaluation (LREC-2006). Genoa, Italy (2006)
5. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: Euzenat, J., Domingue, J. (eds.) AIMSA. Lecture Notes in Computer Science, vol. 4183, pp. 77–86. Springer (2006)
6. Landauer, T.K., Folt, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes 25(2), 259–284 (1998)
7. PeÃśas, A., Rodrigo, l.: A simple measure to assess non-response. In: Lin, D., Matsumoto, Y., Mihalcea, R. (eds.) ACL. pp. 1415–1424. The Association for Computer Linguistics (2011)
8. Witte, E.F.I.H.: Data Mining. Practical Machine Learning Tools and Techniques (2005)