Lexical-Syntactic and Graph-Based Features for Authorship Verification

Notebook for PAN at CLEF 2013

Darnes Vilariño, David Pinto, Helena Gómez
Saúl León, and Esteban Castillo

Benemérita Universidad Autónoma de Puebla
Faculty of Computer Science, Mexico

{darnes, dpinto,saul.leon,helena.adorno}@cs.buap.mx, ecjbuap@gmail.com

Abstract. In this paper we present the results obtained by an approach submitted to the author identification task of PAN 2013 which uses lexical, syntactic and graph-based features for constructing a representation model of document authors. In particular, the features extracted from the graph representation were obtained by means of the SubDue mining tool. As a classification model we have employed Support Vector Machines (SVM). The overall results have ranked our approach in the fifth place from around 17 teams.

Keywords: Authorship verification, graph-based representation, phrase-level lexical-syntactic features, support vector machines

1 Introduction

Authorship verification is the task of determinining if a document has been written by a given author or not. This task is particularly important for forensic linguists who are often called upon to answer this kind of question. This task has been empowered by the continuous growing of information in Internet, thus, the importance of finding the correct features for characterizing the particular writing style of a given author is fundamental for solving the problem of authorship verification.

The results reported in this paper were obtained in the framework of the International Workshop on Plagiarism detection, Author Identification, and Author Profilling (PAN'13). In particular, in the task named "Author Identification" which has focused this year in the problem of authorship verification which may be described as follows:

"Given a small set (no more than 10, 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".

In order to tackle this problem, we propose to extract a set of lexical syntactic level features from each target document, and up to 100 words which are representative of each author. These representative words are selected through the tool "SubDue" (described in Section 2.2) in order to construct a representation of the whole documents written by the given author using a graph structure.

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 SVM classification method employed 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 The proposed approach

In this work we combine two different types of feature extraction. The first one considers lexical-syntactic features, whereas the second uses a data mining based process for extracting the most relevant terms of the target documents. The two features extraction process are described first, and thereafter, we present the manner we construct the final text representation combining the two types of features.

2.1 Lexical-syntactic features

This approach considers the following lexical-syntactic features for representing the particular writing style of a given author:

- Phrase level features
  - Word sufixes. A group of letters added after a word base to alter its meaning and form a new word.
  - Stopwords. A group of words that bear no content or relevant semantics which are filtered out from the texts.
  - Punctuation marks.
  - Trigrams of PoS. Sequences of three PoS tags appearing in the document. Each word text is tagged with its corresponding PoS tag according to the target language. For Spanish language we used the TreeTagger[1], for the English language we employed the Stanford PosTagger[2], whereas for the Greek language we used the Greek POS tagger[3].
- Character level features
  - Vowel combination. Consonants are removed from words and, thereafter, the remaining vowels are combined. Each vowel combination is considered to be a feature. Adjacent repetition of vowels are merged together, considering them as only one vowel.
  - Vowel permutation. Word consonants are removed and, thereafter, the vowel permutation is considered to be a feature.

The text representation schema using the above mentioned features is described in Section 2.3.

[1] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[2] http://nlp.stanford.edu/software/tagger.shtml
[3] http://nlp.cs.aueb.gr/software_and_datasets/AUEB_POS_tagger_2_1_alpha.tar.gz

2.2 Graph-based features

A graph based representation is considered in this approach. Formally, given a graph $G = (V, E, L, f)$ with V being the non-empty set of vertices, $E \subseteq V \times V$ the edges, L the tag set, and $f : E \rightarrow L$, a function that assigns a tag to a pair of associated vertices.

This graph-based representation attempt to capture the sequence among the sentence words, so as the sequence among their PoS tags with the aim of feeding a graph mining tool which may extract relevant features that may be further used for representing the texts. Thus, the set V is constructed from the different words and PoS of the target document.

In order to demonstrate the way we construct the graph for each phrase, consider the following text phrase: "second qualifier long road leading 1998 world cup". The associated graph representation is shown in Figure 1.

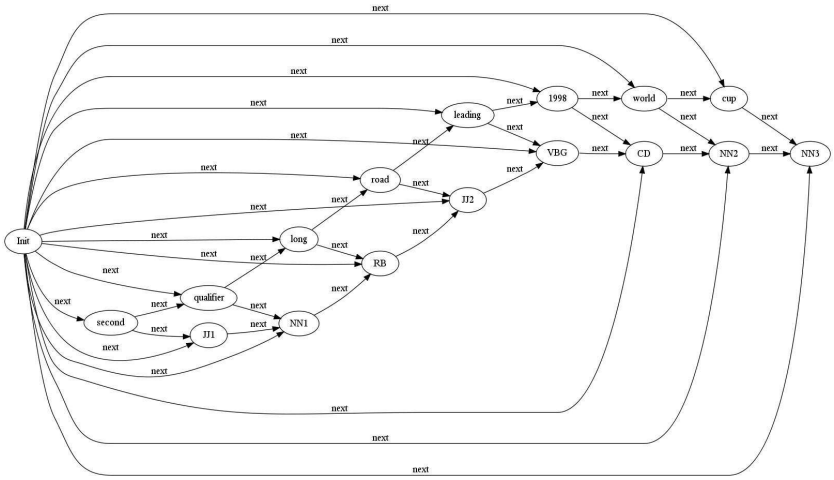

Fig. 1. Graph based text representation with words and their corresponding PoS tags

The process for extracting relevant features from the constructed graph is given as follows.

The Subdue tool Once each paragraph is represented by means of a graph, we apply a data mining algorithm in order to find subgraphs. Subdue is a data mining tool widely used in structured domains. This tool has been used for discovering structured patterns in texts represented by means of graphs [1]. Subdue uses an evaluation model named "Minimum encoding", a technique derived from the minimum description length principle [2], in which the best graph sub-structures are chosen. The best subgraphs are those that minimize the number of bits that represent the graph. In this case, the number of bits is calculated considering the size of the graph adjancency matrix. Thus, the best substructure is the one that minimizes $I(S) + I(G|S)$, where $I(S)$ is the number of bits

required to describe the substructure S, and $I(G|S)$ is the number of bits required to describe graph G after it has been compacted by the substructure S.

2.3 Text representation schema

Let $(x_1, x_2, x_3, \cdots, x_n)$ be the set of features selected for representing the documents, combining the lexical-syntactic and the graph-based features. Each document D is represented considering the feature frequency. Thus, the training stage uses the following feature vector:

$$D = (\underbrace{x_1, x_2, x_3, \ldots, x_n}_{Document\ features}, C) \tag{1}$$

where C is the class manually associated to the document, in this case, the author Name or ID.

For the testing stage, we use the feature vector as follows:

$$D = (\underbrace{x_1, x_2, x_3, \ldots, x_n}_{Document\ features}) \tag{2}$$

In this case, there is not a classification attribute (class name) due to the anonymous source of the document.

3 Description of the classifier used in the task

We have used a Support Vector Machine (SVM) classifier for the task. SVM is a learning method based on the use of a hypothesis space of lineal functions in a higher dimensional space induced by a kernel, in which the hypotheses are trained by one algorithm that uses elements of the generalization theory and taken from the optimization theory.

The linear learning machines are barely used in major real world applications due to their computational limitations. Kernel based representations are an alternative for this problem proyecting the information to a feature space of higher dimensionality which increases the computational capacity of the linear learning machines. The input space X is mapped to a new feature space as follows:

$$x = \{x_1, x_2, \ldots, x_n\} \rightarrow \phi(x) = \{\phi(x)_1, \phi(x)_2, \ldots, \phi(x)_n\} \tag{3}$$

By employing the kernel function, it is not necessary to explicitly calculate the mapping $\phi : X \rightarrow F$ in order to learn in the feature space.

In this research work, we employed as kernel the polynomial mapping, which is a very popular method for modeling non-linear functions:

$$K(x, x) = (\langle x, x \rangle + c)^d \tag{4}$$

where $c \in R$.

For the experiments carried out in this paper, we used the Weka data mining platform[3] for executing the implementation of the SVM classifier.

4 Experimental results

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

4.1 Data sets

For each author, a set of documents of his authorship is given, together with one document which needs to be verified whether or not it has been written by this author. Thus, for our system, all the "known" documents (verified authorship) written by different authors are part of the training set, whereas all the "unknown" documents form the test data. With the training set we generated a model using the SVM classifier, which was further used for classifying the test documents. If the classifier assigns the author that corresponds to each unknown document, the answer is positive, otherwise it is negative.

4.2 Results obtained in the task

The same methodology is applied to the three different languages, considering only some language particularities such as the Greek vowels and the PoS taggers. In Table 1 it can be seen the overall results obtained for each one of the teams that have participated in this edition of the author identification task of PAN 2013. The system proposed by our team (**ayala13**) obtained the fifth place from 17 teams. Given that this approach uses graph mining techniques through the SubDue tool, it can be observed that the runtime is greater than most of the runs submitted to the competition.

Table 1. Overall results for the PAN 2013 authorship verification task

Submission	F_1	Precision	Recall	Runtime
seidman13	0.753	0.753	0.753	65476823
halvani13	0.718	0.718	0.718	8362
layton13	0.671	0.671	0.671	9483
jankowska13	0.659	0.659	0.659	240335
ayala13	**0.659**	**0.659**	**0.659**	**5577420**
bobicev13	0.655	0.663	0.647	1713966
vladimir13	0.612	0.612	0.612	32608
ghaeini13	0.606	0.671	0.553	125655
vandam13	0.600	0.600	0.600	9461
moreau13	0.600	0.600	0.600	7798010
jayapal13	0.576	0.576	0.576	7008
grozea13	0.553	0.553	0.553	406755
gillam13	0.541	0.541	0.541	419495
kern13	0.529	0.529	0.529	624366
baseline	0.500	0.500	0.500	–
petmanson13	0.448	0.700	0.329	20671346
zhenshi13	0.417	0.800	0.282	962598
sorin13	0.331	0.633	0.224	3643942

In Table 2 we present the results obtained by our approach with each one of the three datasets considered in the competition. Three different languages were tackle out. The best performance was obtained with the English language ($F_1 = 0.733$), followed by an $F_1 = 0.667$ in the corpus of Greek documents. However, we obtained a low performance in the Spanish corpus ($F_1 = 0.560$), a result we consider obtained because of the PoS tagger used in the experiments. Further analysis will investigate this issue. It is worth to notice that we always performed better than the competition baselines.

Table 2. Results obtained in different languages

Language	F_1	Precision	Recall	Rank
English	0.733	0.733	0.733	3rd
Greek	0.667	0.667	0.667	3rd
Spanish	0.560	0.560	0.560	8th

5 Conclusions

We have presented an approach that uses two types of features: lexical-syntactic and graph-based. 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 English and Greek languages, but it is a good oportunity for analyzing the text into more deep 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 PoS taggers. As future work, we would like to observe the performance of the proposed methodology using the FreeLing PoS tagger instead of TreeTagger.

When the graph-based features were selected, we empirically determined to extract at most 100 relevant terms using the SubDue graph mining tool. However, more experiments should be performed to analyze whether or not this number introduces significant changes in the obtained results.

References

1. Olmos, I., Gonzalez, J.A., Osorio, M.: Subgraph isomorphism detection using a code based representation. In: FLAIRS Conference. pp. 474–479 (2005)
2. Rissanen, J.: Stochastic Complexity in Statistical Inquiry Theory. World Scientific Publishing Co., Inc., River Edge, NJ, USA (1989)
3. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Series in Data Management Sys, Morgan Kaufmann, second edn. (June 2005)