Lexical-Syntactic and Graph-based Features

for Authorship Verification

Darnes Vilariño, David Pinto, Helena Gómez,

Saúl León, Esteban Castillo

*Benemérita Universidad Autónoma de Puebla, Faculty of Computer Science, Puebla, 72000, Mexico*

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

Discovering the correct features in a raw text which allows unambiguously to attribute the authorship of a given anonymous document is a very hard task. In recent years, there have been a number of research papers in this direction. The traditional authorship attribution task consists in determining the correct authorship of an anonymous document, using a supervised collection of documents, i.e., a reference set of documents manually tagged with their corresponding authorship attribution. In other words, this task can be seen as a classification problem in which the target tag or class is the author name/ID.

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” in order to construct a representation of the whole documents written by the given author using a graph structure.

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 opportunity 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.