Graph-based and Lexical-Syntactic Approaches

for the Authorship Attribution Task

Esteban Castillo1, Darnes Vilariño1, David Pinto1

Iván Olmos1, Jesus A Gonzalez2, Maya Carrillo1

*1Benemérita Universidad Autónoma de Puebla, Faculty of Computer Science, Puebla, 72000, Mexico*

*2 Institute of Astrophysics, Optics, and Electronics, Computer Science Department, Tonantzintla, 72840, Mexico*

*ecjbuap@gmail.com,{darnes, dpinto, iolmos, cmaya}@cs.buap.mx, jagonzalez@inaoep.mx*

Discovering the correct features from a raw text documents collection to unambiguously find the authorship attribution of an 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 of 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.

Determining the authorship of an anonymous document is a task that has been tackled for several years by the computational linguistic community. This is an effort that has been empowered by the continuous growing of information in Internet. 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 problem of authorship attribution.

The results reported in this paper were obtained in the framework of the 6th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN’12). In particular, in the task named “Traditional Authorship Attribution”.

In summary, we presented an evaluation on the use of two different approaches for tackling the authorship attribution task problem. On the first hand, we used a graph to represent text paragraphs by means of words and their corresponding part of speech tags. We aimed to consider the morphosyntactical structure of the text (at once) for further discovering of the best features for the final representation of the training and test datasets. On the other hand, we evaluated a set of six lexical-syntactic features with the purpose of determining those that allow determining an appropriate signature for a given author. A higher number of features (phrase and word level) were independently evaluated, and those that provided the best discrimination scores were selected for the final evaluation.

In general, we observed that the graph-based representation obtained a better performance than the other one. However, more investigation on the graph representation is still required, so that graph patterns discovered by the Subdue tool are better than the ones obtained up to now. As future work, we want to experiment with different graph-based text representations that allow us to obtain much more complex and significative patterns.