An Overview of the Traditional Authorship Attribution   
Subtask  
Extended Abstract

Patrick Juola

Evaluating Variations in Language Laboratory

Duquesne University, Pittsburgh, USA

*juola@mathcs.duq.edu*

Authorship attribution, and specifically the empirical validating and testing of authorship attribution methods, has become an increasingly important and popular research activity. As a followup to our subtask at PAN/CLEF 2011 (Amsterdam), we established a new corpus for analysis for 2012 (Rome). The new corpus differed in several ways from the previous subtask:

* Both the number and size of documents were decreased
* The documents were taken from a different genre (fiction, represented by the *Feedbooks.com* site)
* The documents were no longer marked up extensively
* A new sub-sub-task was added : Authorship clustering. In this new problem (related to “intrinsic plagiarism”) participants were given a text of mixed authorship and asked to determine which paragraphs came from which authors.

The resulting corpus consisted of eight problems, including three closed-class authorship attribution problems, three open-class (the set of correct answers included “none of the above”), and two clustering problems. Twenty-five teams participated in this subtask from many different parts of the world. Detailed results are available on the Web at pan.webis.de and will be discussed in detail at the PAN/CLEF 2012 meeting in September.