**Sub-Profiling by Linguistic Dimensions to Solve the Authorship Attribution Task**

Upendra Sapkota and Thamar Solorio

University of Alabama at Birmingham

Department of Computer and Information Sciences

upendra@uab.edu and solorio@cis.uab.edu

In this paper, we describe a modified version of the profile-basedapproach for the Authorship Attribution (AA) task of the PAN 2012 challenge. In a recent work, Solorio *et al.* (2011) showed that representing documents as a set of separate linguistic modalities in a standard machine learning approach yields good results in AA. On the other hand, profile-based (PB) approaches have shown consistently promising results for the same task. Our PAN system for AA combines these two ideas. We concatenate all the training documents from the same author and build author-specific sub-profiles, one per linguistic modality. Then instead of using all the different types of features to compute the similarity of a test document against an author’s profile in a single step, we compute several similarity scores using one set of features (modality) at a time. We allow each modality to make authorship predictions based on the cosine similarity of the author’s sub-profiles in that modality. Final predictions are based on the combination of decisions from each modality using majority voting. We use stylistic, syntactic, semantic, character *n*-grams and word *n*-grams, resulting in five different modalities. To increase the prediction performance as well as decrease the computation overhead, we decided to perform feature selection based on information gain.

The open class task is much harder than the closed class as test documents may belong to an unknown class. For the open class problem, we have an extra modality containing perplexity values from 4-gram language models at the character level. To determine if a test document belongs to an unknown author, we compare the highest and the second highest cosine similarity score of the author sub-profiles for each modality. If the difference between them is smaller than the threshold value, we decide that the test instance belongs to none of the authors in that modality. We consider the threshold a sort of filter for confusing cases, and we assign those confusing cases to the “unknown" category.

For each dataset in the closed-class problem, our system gave the accuracy which was the highest in PAN 12 challenge. Similarly, the open-class AA framework, although could not perform like the closed-class system, was also very competitive in the challenge.