Author Profiling for English and Spanish Text
Notebook for PAN at CLEF 2013

Upendra Sapkota[1], Thamar Solorio[1], Manuel Montes-y-Gómez[2], and Gabriela Ramírez-de-la-Rosa[1]

[1] University of Alabama at Birmingham, Birmingham, AL 35294, USA
{upendra,solorio,gabyrr}@cis.uab.edu,
[2] Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico
mmontesg@ccc.inaoep.mx

Abstract This paper describes an approach for the author profiling task of the PAN 2013 challenge. This work is based on the idea of linguistic modality[3] that has been successfully used in other classification tasks such as authorship attribution. We consider three different modalities: syntactic, stylistic, and semantic, each representing a different aspect of text. For each modality, we extract informative meta features by computing the similarity relations between the feature vectors in the test files and the centroids of modality specific clusters. Since we were provided texts in both Spanish and English, we build a language independent framework for author profiling. For both English and Spanish documents, our system performed well for the age identification task. For gender prediction, although our system could not perform as expected for English, it yielded good results on Spanish.

Keywords: author profiling, linguistic modality, similarity relations

1 Introduction

With the proliferation of social media (blogs, chats), it has been possible to access large online texts. Such large texts can be utilized to understand how the writing style among users of different age groups, as well as between male and female vary [1,2]. Unlike the authorship attribution problem where the task is to identify the true author of a given piece of text, the author profiling task tries to learn as much information as possible (demographics, personality) about the unknown writer of the given text. Author profiling has a number of application areas such as forensic, security, and commercial domains.

Following the concept of linguistic modalities explored by previous research and proven to be successful for authorship attribution [5,6], we build a framework to solve the author profiling problem. The ultimate target of processing the features separately by linguistic modalities is to compute informative meta characteristics by performing unsupervised clustering of each set of modality vectors in the training set. Solorio *et al.* (2011) mention that the motivation behind using meta characteristics is to leverage

[3] Each linguistic modality refers to a type of feature.

the similarity patterns of each class along the different linguistic modalities in the form of meaningful higher level features. In this paper, we explore the idea of linguistic modalities for author profiling task. We also explore various features to understand their contribution in the author profiling task.

To automatically identify the age and gender of the unknown author of a given text either in English or Spanish, we consider only the language independent features on different modalities. For both languages, we use the same features that help us understand how the performance of the author profiling system changes with language.

2 Methodology

Each document is represented separately by modality, i.e., a document will have one feature vector per modality. We denote this set of vectors as First Level Features (FLF) following the naming convention from [6]. Once FLF for all the training documents are extracted, the first step towards meta feature extraction is modality-wise clustering of the feature vectors of training documents. This results in k clusters for each modality. Modality Specific Meta Features (MSMF) are the cosine similarities between a test feature vector and centroid of each cluster (average of feature vectors in the cluster). If we consider a total of m modalities, and if each modality generates k clusters, we have a total of $m \times k$ meta features.

The outline of the algorithm is given below.
1. For each modality
1.1 Cluster all the training vectors in an unsupervised fashion to get k clusters.
1.2 Compute the cluster centroid by averaging the feature vectors in that cluster.
1.3 Generate meta feature by computing cosine similarity [4] between cluster centroid and each (both test and train) document's feature vector.

For the author profiling problem, we consider three different modalities –stylistic, semantic, and perplexity values from character n-gram language models. In the end, we augment meta features with FLF to train a machine learning classifier.

3 Experimental Results and Evaluation

For all our experiments, the semantic modality contains the top 5000 words, while the perplexity modality contains a perplexity value for each class computed from character 4-gram language models. In the stylistic modality, we have 22 features. In this modality, 12 features are taken from [5] and we added 10 new features, most of which are related to HTML format as well as errors regarding the use of the indefinite articles. Based on preliminary results, we added 1) HTML features such as count of 'www', count of '<img', count of '<a href', and count of '<br'; 2) emoticon based features such as count of happy emoticons as well as count of all emoticons; 3) count of mistakes in the use of 'a' and 'an'. To compute these counts, we use several regular expressions. Besides the addition of these features, we also investigated the use of other types of features as new modalities. We experimented to see if the expression of certain type of emotions

in the text could help in the profiling task. For this, we used a number of emotion-based features as a new modality but preliminary results showed no improvement in the performance, and hence we discarded these features. Similarly, we experimented with different features extracted from POS tagged data, but to our surprise, we observed degradation in the performance. In the semantic modality, we experimented the use of tf–idf weighting, but then also there was no improvement in the performance. We finally decided to use simple normalized term frequency to create the feature vectors. We performed all these experiments using only 15,000 documents from the total training data and this might have caused us to make decisions that ended up hurting performance on the test set.

Documents in two languages: English and Spanish were considered for the profiling task. There are two profiling problems: determining the author's gender (Male and Female) and author's age (10s: 13-17, 20s: 23-27 and 30s: 33-47), which we solved as a single six class problem.

In the final submission, only in the stylistic and perplexity modalities, we included both the FLF and meta features while the semantic modality includes only the meta features. We obtained a largely reduced feature set containing only 208 features in total. The independent features from different modalities at the end are merged together to create a single feature vector per document, which is the final document representation. Accuracy was the only performance measure for evaluating the software. We used support vector machines (SVMs) implemented in Weka [8] with default parameters as the underlying classifier. To train the character 4-gram language models, we applied the SRILM toolkit [7]. The clustering was done using CLUTO [3]'s $vculster$ clustering program with parameter $clmethod = rbr$ that selects a k-ways clustering solution, which at the end is globally optimized .

Table 1. Performance of our system on both languages. The last row is baseline performance.

Method	**Language**	**Accuracy**		
		Total	Gender	Age
Our PAN13 system	English	0.2471	0.4781	0.5415
	Spanish	0.2934	0.5116	0.5651
Baseline	both	0.1650	0.5000	0.3333

For both English and Spanish, our author profiling framework performed well for age prediction, yielding accuracy of $\approx 55\%$, that is ≈ 1.7 times better than the baseline. We did not obtain similar performance on the gender identification, although for Spanish, our system is slightly better than the baseline. When we consider identification of age and gender as a single compound task, accuracy of our system for English is 24.71%, ≈ 1.5 times better than the baseline, and for Spanish is 29.34%, that is ≈ 1.8 times better than the baseline. For this compound task (age + gender), the best performance of the PAN 2013 challenge was also not that high, 38.94% for English and 42.08% for Spanish. This may be due to the complexity of the task. In Table 1, although the average performance of our system for age identification is 55.33% and 49.46% for

gender identification, the overall average accuracy is much lower (27.03%). However, when we average the accuracy of English and Spanish, our system is still ≈ 1.63 times more accurate than the baseline.

In this project, we explored different features for author profiling and tried to understand whether a framework used for other classification task, such as authorship attribution could be adapted to author profiling. After various experiments, we believe that we can improve results on author profiling by including some task-specific features.

4 Conclusions

In this paper, we showed how a modality specific approach used successfully for AA would perform on the profiling task. We solved two profiling problems: gender and age as a single six class problem. Our method uses only 208 features and is competitive, yielding acceptable accuracies on both profiling problems on the Spanish dataset. However, for English, our system gave poor performance, specially for gender identification. But still, we were able to have a fully automatic language-independent author profiling system that on average performs ≈ 1.63 times better than a simple baseline. A good future direction would be to evaluate our approach on different datasets as well as to explore the task-specific features to improve the performance.

Acknowledgements

This research was partially supported by ONR grant N00014-12-1-0217 and by NSF award 1254108. It was also supported in part by the CONACYT grant 134186.

References

1. Argamon, S., Dhawle, S., Koppel, M., Pennebaker, J.W.: Lexical predictors of personality type. In: Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America (2005)
2. Corney, M., de Vel, O., Anderson, A., Mohay, G.: Gender-preferential text mining of e-mail discourse. In: Proceedings of the 18th Annual Computer Security Applications Conference. pp. 282–. ACSAC '02, IEEE Computer Society, Washington, DC, USA (2002)
3. Karypis, G.: CLUTO - a clustering toolkit. Tech. Rep. #02-017 (Nov 2003)
4. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing Management 24(5), 513 – 523 (1988)
5. Sapkota, U., Solorio, T., Montes-y Gómez, M., Rosso, P.: The use of orthogonal similarity relations in the prediction of authorship. In: Computational Linguistics and Intelligent Text Processing. pp. 463–475. Springer Berlin Heidelberg (2013)
6. Solorio, T., Pillay, S., Raghavan, S., Montes-y Gómez, M.: Generating metafeatures for authorship attribution on web forum posts. In: Proceedings of the 5th International Joint Conference on Natural Language Processing, IJCNLP 2011. pp. 156–164. AFNLP, Chiang Mai, Thailand (November 2011)
7. Stolcke, A.: SRILM - an extensible language modeling toolkit. pp. 901–904 (2002)
8. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kauffmann, 2nd edn. (2005)