Semantic-based Features for Author Profiling Identification: First insights

Delia-Irazú Hernández[1], Rafael Guzmán-Cabrera[2], Antonio Reyes[3] and Martha-Alicia Rocha[1,4]

[1]Universidad Politécnica de Valencia, España
[2]Universidad de Guanajuato, México
[3]Instituto Superior de Intérpretes y Traductores, México
[4]Instituto Tecnológico de León, México

Introduction

Nowadays, social interaction through Internet is becoming a major problem due to the insufficient control regarding the authenticity of user profiles. Author profiling is the task of identifying personal characteristics of Internet users (such as age, gender, native language) based on analysing their interactions, mainly, considering their textual patterns. Author profiling has various applications such as security, forensics, marketing, among others.

Features Description

- **Signature** $\rightarrowtail$ explicit linguistic markers within a text.
- **Chatslang** $\rightarrowtail$ words and expressions commonly used in internet forums.
- **Context** $\rightarrowtail$ the presence of discriminating clusters across the classes.
- **Emotionality** $\rightarrowtail$ the use of words to communicate emotions, feelings, moods, etc.
- **Semantic similarity** $\rightarrowtail$ the semantic relatedness among the words of a text.
- In addition a list of **Bag of Words** was used. Finally, the **Jaccard similarity** coefficient was applied in order to focus on informative words rather than only on frequent ones.

Experiments

- Each conversation is represented as a numerical vector in which each entry represent a feature.
- We make different combinations of the features proposed and we classified the conversations using various learning algorithms.

Experiment	Description
SBF	Semantic similarity + Signatures + ChatSlang + Emotionality
SBF_M	Semantic similarity + Emotionality
SBF+BOW	***SBF*** + Bag of Words
SBF+Jaccard	***SBF*** + Jaccard similarity coefficient
SBF+Jaccard+BOW	***SBF+BOW*** + Jaccard similarity coefficient
Jaccard	Jaccard similarity coefficient
Jaccard+BOW	Jaccard similarity coefficient + Bag of Words
Jaccard+Context	Jaccard similarity coefficient + Context

Table: Features Combination

Results

The learning algorithms applied to this combinations were Naive Bayes (NB), Support Vector Machines (SVM), Multilayer Perceptron (MP), Decision Tree (J48), and a bagging of classifiers (NB + SVM +J48). The table below introduces the results obtained from different experiments.

Experiments	NB	SVM	MP	J48	Bagging	Average
SBF	16.66	19.66	18.3	17.66	18.67	17.99
SBF+BOW	22	15.66	-	21.66	20.67	19.77
SBF+Jaccard	19	20.60	18.33	15.33	20	18.31
SBF+Jaccard+BOW	23	15.66	-	22	21	**20.22**
Jaccard	22.33	17.33	22.33	14.67	17.66	18.11
Jaccard+BOW	23	15.33	-	19.33	21.33	19.22
Jaccard+Context	17.66	21.6	-	17.33	21.33	18.86

Table: Results of Author Profiling classifiers

PAN results

- We defined a final model which was integrated with the features: ***SBF_M*+Jaccard+BOW**. According to our best results, the NB classifier was used to participate in the author profiling task in PAN 2013 competition. The results are shown in the next table.

Task	Accuracy		
	total	Gender	Age
English	0.2816	0.5671	0.5061
Spanish	0.1757	0.4982	0.3554

Table: Author Profiling Evaluation PAN 2013

Conclutions

- This research is based on semantic features.
- After analyzing the results, we could realize that the author profiling task has a high lever of overlap between classes; hence, the dificulty of correctly identifiying the classes increases subtiantially.
- The future work consists of developing an algorithm for principal components analyisis (PCA) in order to obtain highly discriminating features.