Two methodologies applied to the author profiling task

Yuridiana Aleman, Nahun Loya, Darnes Vilariño, David Pinto

Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla, México
candy.aleman@cs.buap.mx, nahun.loya@cs.buap.mx, darnes@cs.buap.mx,
dpinto@cs.buap.mx

Abstract. This paper describes two methodologies applied to the author profiling task submitted to the PAN 2013 competition of the CLEF 2013 conference. The first methodology was applied only to the English language, whereas the second one was executed only over the corpus written in Spanish language. The aim was to evaluate the performance of both methodologies in the above mentioned task. The obtained results were quite positive for the first methodology which considers a classicaly approach of classification, using diverse features extracted from the texts in order to feed a classifier based on random forests. The second methodology, based on graph mining techniques, obtained a very poor performance for the author profiling task.

1 Description of the Methodologies Evaluated

We applied two different methodologies, one for each language. For the English corpus, we employed machine learning techniques with different sets of features. The description of this first methodology is presented in Section 1.1. The Spanish corpus was processed with a second methodology based on graph mining techniques. This methodology is described in Section 1.2.

1.1 English

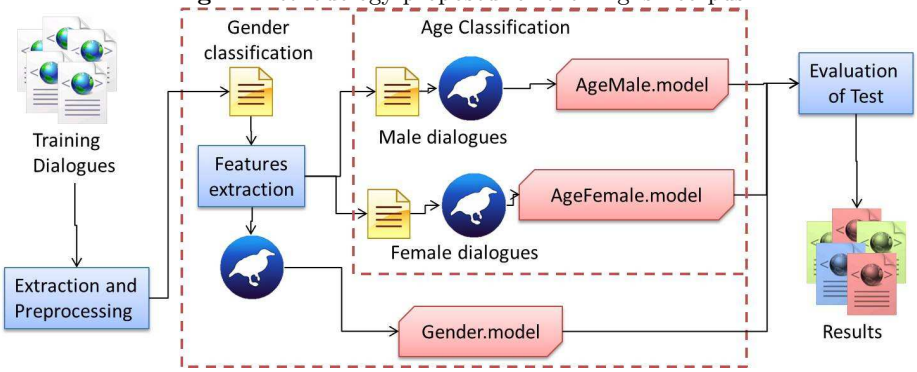
For the English corpus, we applied a methodology based in classical techniques of machine learning. The set of features were extracted in order to feed a Random Forest classifier. Figure 1 shows the methodology used for this corpus which is twofold: pre-processing and classification.

In the pre-processing step, we attempt to normalize terminology by replacing unrecognizable terms, smiles, and weird symbols (e.g. URLs, pictures) from the dialogues by their corresponding normalized term. In order to apply this normalization procedure, we used three lexical resources that we have constructed for this purpose. The lexical resources are described as follows:

1. **Emoticons**: A list of emoticons constructed on the basis of a preliminar dictionary[1], and enriched by adding the predefined emoticons of *Windows*

[1] http://netlingo.com/smileys.php

Fig. 1. Methodology proposed for the English corpus



Messenger, *Facebook* and *Gmail*. The dictionary constructed contains 344 entries.

2. **Contractions**: The list contains around 65 of the most used contractions in the United States.
3. **Dictionary**: A English dictionary in TXT format. This list was used for determining unrecognized words.

All the term occurrences in the dialogues of the training and test set that matches with some entry in "emoticons" or "Contractions" were replaced. The "dictionary" lexical resource was used only for determine the existence of a given term of the conversation.

In the classification process we used the frequencies of the following sets of features:

- Emoticons
- Contractions
- Conversation length (in words)
- Conversation length (in characters)
- Mispelled words
- Average length of words in the dialogues
- Words capitalized
- Words in uppercase
- URLs
- Each different POS tag
- Each different suffix
- Each different punctuation symbol
- Each stopword

All these features were used for representing each one of the dialogues in the training set which further were used for feeding a *Random Forest* classifier (included in the *WEKA* tool[1]). The gender was used as the classifier attribute

(class) for determining whether a given dialogue was written by a male or a female person.

A second classification model consider the discrimation of the age. In this case, the classifier attribute is the range of age ("10s", "20s" or "30s") given at the competition. The result of the first classifier (gender classifier) determines which type of the second classifier will be used, the one that was trained only with a corpus of male persons, or the one that was trained only with female persons. In summary, we obtained three classification models:

1. Classification by gender (*Gender.model*).
2. Classification of age range for male persons (*AgeMale.model*).
3. Classification of age range for female persons (*AgeFemale.model*).

We extracted the sets of features in all documents and three models was create, *Gender.model*, *AgeMale.model* and *AgeFemale.model*. The system for the classification of the test dataset takes into consideration the following steps:

1. The extracted dialogues of the test corpus were preprocessed using the lexical resources in order to normalize the texts.
2. We obtained the set of features for the gender classifier. Afterwards, we classified the test dialogues for obtaining a gender label for each dialogue ("male" or "female").
3. We separated the dialogues according to the gender label, then, we classified the female dialogues with *AgeFemale.model*, and the male dialogues using *AgeMale.model*.
4. Finally, each test dialogue has two categories assigned, age and gender. Using these categories, the system prepare the sytem output in XML format.

The results obtained in the competition are given in Section 2.

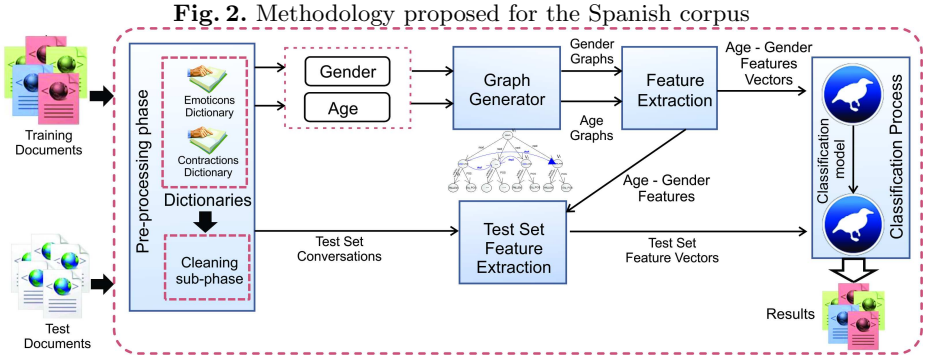
1.2 Spanish

The methodology proposed for the Spanish corpus focuses on the use of graphs as a strategy for feature extraction. Moreover, this methodology uses the extracted features with the purpose of feeding a supervised classification algorithm which allows to determine the gender and age of the authors. As carried out with the English corpus, we performed a pre-processing step in order to normalize the input texts. Afterwards, the texts are represented by means of graphs which will be further used for extracting relevant features. Graphs are mined using the SUBDUE tool[2]. The obtained results of the mining graph phase are used as features in machine learning algorithms with the aim to obtain a classification model.

A graphic idea of the solution scheme is shown in Figure 2.

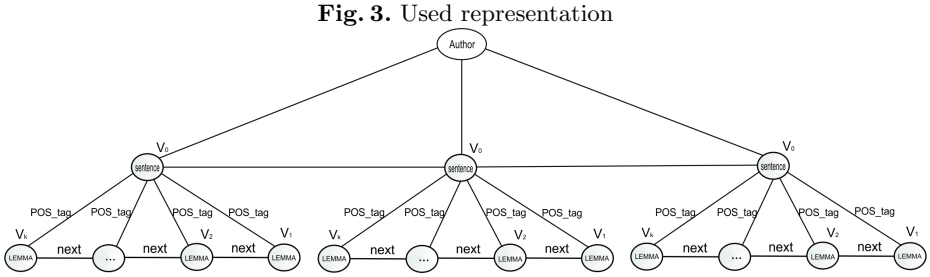
[2] `http://ailab.wsu.edu/subdue/`

Fig. 2. Methodology proposed for the Spanish corpus



Graph Generator The Graph Generator module receives the training set split into two subsets, one made up of the dialogues labeled with age classes, and the other one labeled with the gender classes. A graph-based representation for each class is built using a star topology. The star topology allows depict text in sentences. This graph-based representation generates a vertex in the graph for each word in the sentence. Furthermore, the representation establishes the relationships between words through of the edges of the graph. A graphical scheme is shown in the figure 3

Fig. 3. Used representation



Before the generation process begins all dialogues are tagged used the freeling[3] tool with the aim of finding the word lemmas and the Part-of-Speech tags (POS tags). Thereafter, an order between words is established; that given order is symbolized by a generic conector "next". The process begins establishing an inicial vertex v_0 representing a sentence, from which adjacent vertices are derived v_1, v_2, ..., v_k, where each one represents a lemmatized word, and k represents the number of words in the sentence. Each edge having an inicial node v_0 and

[3] `http://nlp.lsi.upc.edu/freeling/`

a final node v_k, with $k = 1, \ldots, no\_words$, is assigned with a "POS_tag" representing the POS tag obtained with Freeling. Finally each edge having an initial node v_i and final node $v_i + 1$, with $i = 1, \ldots, k - 1$ is assigned with the generic conector "next".

This module generates a graph-based representation for each class of the gender set (male-female) and the age set (10s,20s,30s).

Feature Extraction The feature extraction is conducted over all graph-based representations; the first one over the gender set (male and female), whereas the second on is conducted over the age set (10s, 20s, and 30s). Finally, based on the analysis of the extracted features of those two sets, a feature set is built for the combined class (10s-male, 10s-female,20s-female,20s-female,30s-female,30s-female).

The feature extraction process over all graph-based representation is done by mining those graphs with the SUBDUE tool. The result of the mining process is analyzed as follows: data containing the highest support values are considered, that is, the substructures found from the graph mining process that frequently appear are used. For this particular step, the MDL measure of the mining tool is used. The results of each mining process for the age set and the gender set are single words or n-grams of words that exclusively appear in one class of the set, but not in the other. For example, words or n-gram that appears in the male set, but no appearing in the female set and viceversa. The feature vector obtained is later used by a supervised classifier.

The test set feature extraction is performed based on the features obtained with the train set, i.e., every feature in the train set is extracted and counted in the test set, generating a feature set with an unknown class.

Classification Process This module receives the train feature vectors with its corresponding class obtained in the previous module, i.e. $D = (C_1, C_2, \ldots, C_n)$, where C_i represent a particular characteristic and n is the total number of the characteristics. We have used the random forest classification algorithm, included in Weka[4] tool in order to obtain a classification model for the six classes (10s-male, 10s-female,20s-female,20s-female,30s-female,30s-female). Thereafter, the test Set feature vectors are evaluated with the generated model, and every vector is labeled with one age-gender class.

2 Experimental results

In this section we present the results obtained with the two proposed methodologies. First we describe the dataset used in the experiments, and thereafter, the results for each language.

[4] `http://www.cs.waikato.ac.nz/ml/weka/`

2.1 Dataset

The English training corpus contains 236,600 dialogues of different types of authors. The corpus is balanced with 118,300 dialogues per gender. Table 1 shows some characteristics of this corpus.

Table 1. English training dataset characteristics

Data	Male	Female
Total of dialogues	118,300	118,300
Dialogues of 10s	8,600	8,600
Dialogues of 20s	42,900	42,900
Dialogues of 30s	66,800	66,800
Vocabulary	410,425	403,539
Words average	715	807
Largest dialogue	18,483	11,037

The male gender part of this corpus contains a larger vocabulary than the female one, however, the female part of this corpus is more extensive in average of words per dialogue.

The Spanish training corpus is structured as shown in the table 2. The number of dialogues is much more smaller than the English one.

Table 2. Spanish training dataset characteristics

Age group	Gender	No. Authors
10s	Male	1,250
	Female	1,250
20s	Male	21,300
	Female	21,300
30s	Male	15,400
	Female	15,400

In this edition of the PAN competition, we were required to submit the entire system to a virtual machine in which the execution of the experiments will be carried out. We did not have access to the test dataset, therefore, the description of this corpus is not given in this paper.

2.2 Obtained results

Tables 3 and 4 show the results obtained at the competition for the English and Spanish corpus, respectively. As already mentioned, the first methodology was

only applied to the English corpus. In this case, we can see that the Accuracy obtained is 0.5923 which rank the system in the 7th position. However, the second methodology performed even worse than the baseline, with an Accuracy of 0.2915. As future work, we would like to evaluate the first methodology in the Spanish corpus as well in order to determine its performance with the Spanish language.

Table 3. Performances on the English portion of the test data

Submission	**Accuracy**			**Runtime**
	Total	Gender	Age	(incl. Spanish)
meina13	0.3894	0.5921	0.6491	383821541
pastor13	0.3813	0.5690	0.6572	2298561
mechti13	0.3677	0.5816	0.5897	1018000000
santosh13	0.3508	0.5652	0.6408	17511633
yong13	0.3488	0.5671	0.6098	577144695
ladra13	0.3420	0.5608	0.6118	1729618
ayala13	0.3292	0.5522	0.5923	23612726
gillam13	0.3268	0.5410	0.6031	615347
kern13	0.3115	0.5267	0.5690	18285830
haro13	0.3114	0.5456	0.5966	9559554
aditya13	0.2843	0.5000	0.6055	3734665
hidalgo13	0.2840	0.5000	0.5679	3241899
farias13	0.2816	0.5671	0.5061	24558035
jankowska13	0.2814	0.5381	0.4738	16761536
flekova13	0.2785	0.5343	0.5287	18476373
ramirez13	0.2471	0.4781	0.5415	64350734
jimenez13	0.2450	0.4998	0.4885	3940310
moreau13	0.2395	0.4941	0.4824	448406705
baseline	0.1650	0.5000	0.3333	–
patra13	0.1574	0.5683	0.2895	22914419
cagnina13	0.0741	0.5040	0.1234	855252000

3 Conclusions and future work

We have presented two different methodologies for tackling out the author profiling task. The main difference between the two approaches is the feature extraction process. The first approach uses a number of features which we consider to be related to the two classes to be discriminated (age and gender). The second approach is based only on the word frequencies and POS tags, but uses a graph-based representation for extracting n-grams of words.

We succeed in the first approach obtaining the 7th place in the competition, but the second one was not able to capture regularities or patterns from the graphs. The results obtained by the first approach indicates that the features selected allow to discriminate gender and age of a given author with an F-measure of 0.59. We are interesting in evaluating the first methodology presented in the Spanish language.

Table 4. Performances on the Spanish portion of the test data

Submission	**Accuracy**			**Runtime**
	Total	Gender	Age	(incl. English)
santosh13	0.4208	0.6473	0.6430	17511633
pastor13	0.4158	0.6299	0.6558	2298561
haro13	0.3897	0.6165	0.6219	9559554
flekova13	0.3683	0.6103	0.5966	18476373
ladra13	0.3523	0.6138	0.5727	1729618
jimenez13	0.3145	0.5627	0.5429	3940310
kern13	0.3134	0.5706	0.5375	18285830
yong13	0.3120	0.5468	0.5705	577144695
ramirez13	0.2934	0.5116	0.5651	64350734
aditya13	0.2824	0.5000	0.5643	3734665
jankowska13	0.2592	0.5846	0.4276	16761536
meina13	0.2549	0.5287	0.4930	383821541
gillam13	0.2543	0.4784	0.5377	615347
moreau13	0.2539	0.4967	0.5049	448406705
cagnina13	0.2339	0.5516	0.4148	855252000
hidalgo13	0.2000	0.5000	0.4000	3241899
farias13	0.1757	0.4982	0.3554	24558035
baseline	0.1650	0.5000	0.3333	–
ayala13	0.1638	0.5526	0.2915	23612726
mechti13	0.0287	0.5455	0.0512	1018000000

References

1. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11**(1) (November 2009) 10–18