PAN2013

Two methodologies applied to the author profiling task

Yuridiana Alemán, Nahun Loya, Darnes Vilariño, David Pinto
Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla
candy.aleman@cs.buap.mx, nahun.loya@cs.buap.mx, darnes@cs.buap.mx, dpinto@cs.buap.mx

We present two methodologies applied to the autor profiling task. The first methodology was applied only to the English language, we used diverse features extracted from the texts in order to feed a classifier based on random forests. The obtained results were quite positive. The second proposal was executed only over the corpus written in Spanish language. It is based on graph mining techniques, obtained a very por performance for the competition .

ENGLISH METHODOLOGY

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.

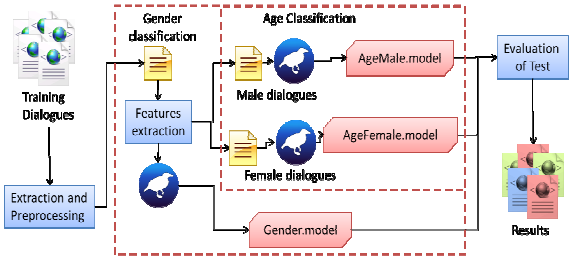

Fig. 1. Methodology proposed for the English corpus

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 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. We obtained three classification models:

1. Classification by gender.
2. Classification of age range for male persons.
3. Classification of age range for female persons

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.
2. We classified the test dialogues for obtaining a gender label for each dialogue.
3. We separated the dialogues according to the gender label, then, we classified the dialogues according with age label.
4. Finally, each test dialogue has two categories assigned, age and gender. Using these categories, the system prepare the sytem output in XML format.

SPANISH METHODOLOGY

For the Spanish corpus, we perform a methodology based in Graph-based representation with the aim to extract features for the classification task. Figure 2 shows the used methodology which consist of four steps: first the preprocessing phase, second the Graph generator, thrid the feature extration based on Graph data mining phase and finally the classification process.

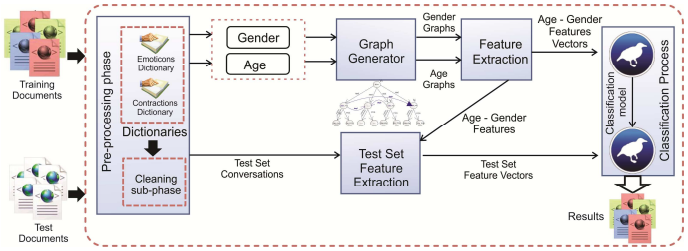

Fig. 2. Methodology proposed for the Spanish corpus

The first step consist in preprocess the text, repacing the emoticons using a dctionary, next is used a dictionary of contractions, finally the texts are cleaned erasing different features like:

- Emoticons
- Contractions
- Stopwords

In the next step we transform the texts to the graphs using a star topology. Then is generated a set of graphs which is are mined to find patterns using a SUBDUE tool.

The output of the Graph dataming tool is analyzed take in account groups of words that appear in the graphs. Similarly to the n-grams model.

Finally the obtanied features are collocated like characteristics on supervised classificators.

We obtained three classifications models based on graphs features.

1. Classificatiion by gender
2. Clasifficaction by age
3. Classification by age-gender

Finally the test data are tested using the features of the age-gender, utilizing the Vote algorithm provided by Weka tool.

The prediction output is transform in a XML format acording to the autor´s number.

RESULTS

The graphs show the results obtained at the competition. As already mentioned, the first methodology was only applied to the English corpus. In this case, we can see that the Accuracy
obtained is 0.33 which rank the system in the 7th position. However, the second methodology performed even worse than the baseline, with an Accuracy of 0.16.