CNG text classification for authorship profiling task

Notebook for PAN at CLEF 2013

Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios

Faculty of Computer Science, Dalhousie University
jankowsk, vlado, eem@cs.dal.ca

Abstract We describe our participation in the Author Profiling task of the PAN 2013 competition. The task objective is to determine the age and the gender of an author of a document. We applied the Common N-Gram (CNG) classifier (Kešelj et al., 2003) to this task. The CNG classifier uses a dissimilarity measure based on the differences in the frequencies of the character n-grams that are most common in the considered documents. To train the classifier, a class is represented by one class document created by concatenating the training documents belonging to the class. A sample document is labelled by the class with the minimum dissimilarity. For the six class classification (combinations of two possible gender labels and three possible age labels) we achieved the accuracy of 0.2814 on the English test dataset and 0.2592 on the Spanish test dataset. Our results are below the medians of the results of the competition participants.

1 Introduction

Author profiling problem is a problem of determining some characteristics of an author of a document, such as demographics. The Author Profiling task of the PAN 2013 competition presents the problem of determining the age and the gender of authors. We tested applying the Common N-Gram (CNG) classifiers proposed by Kešelj et al. [3] to this task.

2 Methodology

The Common N-Gram (CNG) classifier is based on the the differences in the usage frequencies of the most common character n-grams of the considered documents. Given training documents for a given class, the classifier concatenates them into one class document. A sample document to be classified is then compared to each of the class documents using the CNG dissimilarity measure and the sample is labelled by the class with the minimum dissimilarity. To calculate the dissimilarity, for each document a sequence of the most common character n-grams coupled with their frequencies (normalized by the length of the document) is extracted; such a sequence is called a *profile* of the document. The dissimilarity between two documents of the profiles P_1 and P_2 is defined as follows:

$$D(P_1, P_2) = \sum_{x \in (P_1 \cup P_2)} \left(\frac{f_{P_1}(x) - f_{P_2}(x)}{\frac{f_{P_1}(x) + f_{P_2}(x)}{2}} \right)^2 \qquad (1)$$

where x is a character n-gram from the union of two profiles, and $f_{P_i}(x)$ is the normalized frequency of the n-gram x in the profile P_i, $i = 1, 2$ ($f_{P_i}(x) = 0$ whenever x does not appear in the profile P_i).

The important parameters of the dissimilarity is the length of the character n-grams n and the length of the profile L.

The CNG classifier (or its variants) has been successfully applied to the authorship classification tasks [3], [1], [4].

Using the PAN 2013 Author Profiling training corpus provided by the competition organizers we explored how the parameters of the classifier affect the accuracy of the classification. The training corpus consists of XML documents with the HTML content of online conversations. There is English and Spanish subset of the corpus. Each document is labelled by one of two gender classes: male or female, and one of three age classes: “10s”, “20s” or “30s”. The combination of these labels yields six distinct classes for each language. We used part of the training corpus for each class to create the class document for each class (90% randomly selected documents for each class with the exception of English “20s” and “30s” classes — both male and female; these are the largest classes and for the performance reasons we used for the training 20% of documents from the English “20s” classes and 10% from the English “30s” classes).

In our software we used n-grams in which tokens are utf8-encoded characters. The package Text::Ngrams [2] was used to extract the n-grams and their frequencies.

In our experiments (with a balanced subset of the remaining documents from the corpus as test data) using the HTML format of the conversation led to better results than converting the conversations to text format (which was not suprising, as then the common character n-grams can capture such features as line breaks and links). Also one CNG classifier for six classes tended to yield higher accuracy than two separate, independently trained CNG classifiers: one for the gender classification (with two classes) and one for the age classification (with three classes).

Based on our experiments we arrived at the parameters reported in Table1 that we used in the software submitted for the competition.

	English	Spanish
n (n-gram length)	4	5
L (profile length)	5000	5000
format of the data	html	

Table 1. The parameters of the CNG classifier used in the competition.

3 Results

In the PAN 2013 competition task Author Profiling our method yielded the results presented in Table 2.

Total accuracy	Gender accuracy	Age accuracy	Competition rank
English test data			
0.2814	0.5381	0.4738	14th of 21
Spanish test data			
0.2592	0.5846	0.4276	11th of 20

Table 2. The results in the PAN 2013 competition task Author Profiling, according to the results announced on June 12, 2013.

References

1. Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1(3), 233–334 (Dec 2006)
2. Kešelj, V.: Perl Package Text::Ngrams. http://www.cs.dal.ca/ vlado/srcperl/Ngrams (accessed on Feb 1, 2012)
3. Kešelj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03. pp. 255–264. Dalhousie University, Halifax, Nova Scotia, Canada (August 2003)
4. Stamatatos, E.: Author identification using imbalanced and limited training texts. In: Proceeding of the 18th International Workshop on Database and Expert Systems Applications, DEXA'07. pp. 237–241 (September 2007)