CNG text classification for authorship profiling task

Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios

Faculty of Computer Science, Dalhousie University

Introduction

Author profiling problem:
Determining some characteristics of an author of a text, shared by groups of persons, such as age, gender, native language, personality traits, etc.

Applications:

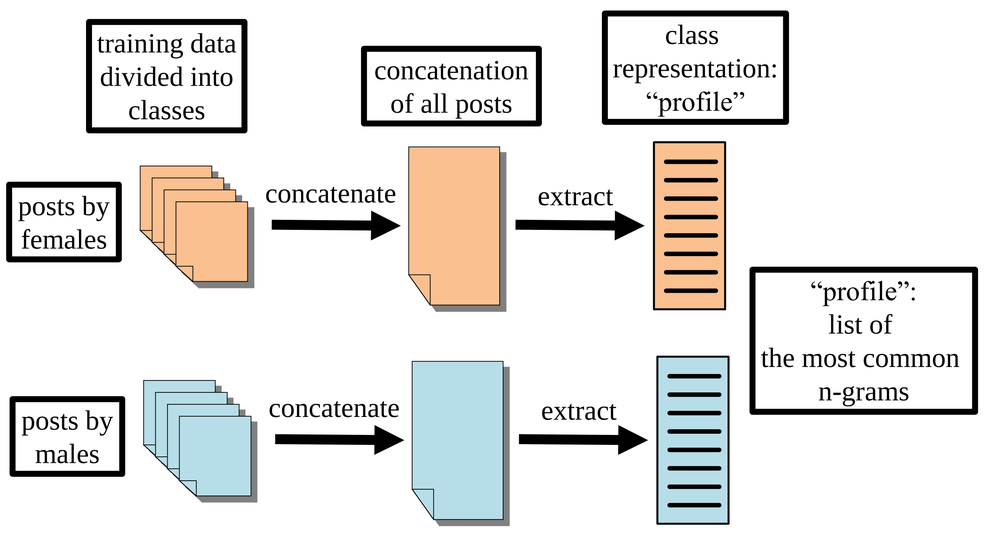
- Forensics
  e.g., gender of an author of a message as an evidence
- Marketing
  e.g., demographics of authors of positive reviews of a product

We tested the use of **Common N-Gram (CNG) classifier** (Kešelj et al., 2003) with character n-grams at the **PAN 2013 Author Profiling competition**.

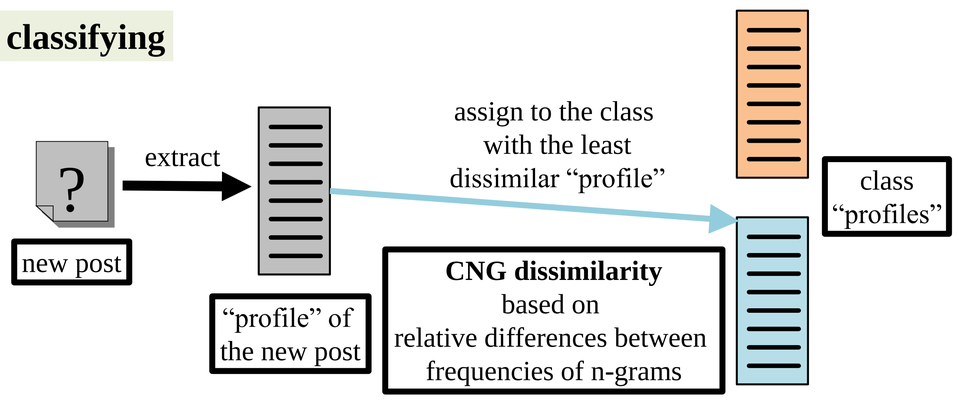
Methodology: algorithm

We applied **Common N-Gram (CNG) Classifier**, proposed by Kešelj et al. [3], which is an effective approach for **authorship attribution** [3,4]. It has been also successfully applied to **author gender classification** for English essays [2].

training



classifying



Methodology: features

We used **character n-grams**: strings of n consecutive characters from a given text.

- language independent
- easy to use even for languages with nontrivial word segmentation (e.g., Chinese, Japanese)
- effective feature for authorship attribution

Results

PAN 2013 Author Profiling competition

- Corpus: blog posts
- Task: age and gender of authors: 6 classes in total
  three age classes: year ranges (13-17), (23-27) , (33-47)
  two gender classes
- English and Spanish datasets

Our competition submission:

- for each language one CNG classifier with 6 class profiles (trained on subsets of PAN 2013 Author Profiling training data)
- no preprocessing: html format of posts
- n-grams of utf8-encoded characters
- length of profiles: 5000 most common n-grams
- length of n-grams: n=4 for English, n=5 for Spanish

Results on PAN 2013 Author Profiling test dataset

	English			Spanish		
	total	gender	age	total	gender	age
accuracy of our CNG method	0.2814	0.5381	0.4738	0.2592	0.5846	0.4276
competition rank	14th of 21			11th of 20		
best accuracy by other participants	0.3894	0.5921	0.6572	0.4208	0.6473	0.6558
random baseline	0.1650	0.5000	0.3333	0.1650	0.5000	0.3333

Conclusions and Future Work

The CNG classifier with character n-grams yielded results below median of all participants accuracies, and within the third quarter of all the participant results.

Combining this n-gram approach with other features for author profiling would be worth investigation, as would be exploring the use of visual analysis for gaining insight and interacting with the classification process for author profiling task.

Bibliography

1. Argamon, S., Koppel, M., Pennebaker, J., Schler, J.: Automatically profiling the author of an anonymous text, Communications of the ACM 52 (2), pp. 119–123 (2009),
2. Doyle, J., Kešelj V.: Automatic categorization of author gender via n-gram analysis. In Proc. of the 6th Symposium on Natural Language Processing, SNLP'2005, (December 2005).
3. Kešelj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proc. of the Conference Pacific Association for Computational Linguistics, PACLING'03. pp. 255–264. (August 2003)
4. Stamatatos, E.: Author identification using imbalanced and limited training texts. In: Proc. of the 18th International Workshop on Database and Expert Systems Applications, DEXA'07. pp. 237–241 (September 2007)

Acknowledgment

This research was funded by a contract from the Boeing Company, a Collaborative Research and Development grant from the Natural Sciences and Engineering Research Council of Canada, and Killam Predoctoral Scholarship.

Contact

{jankowsk, vlado, eem} @cs.dal.ca

Faculty of Computer Science
Dalhousie University
6050 University Avenue
PO BOX 15000
Halifax, NS B3H 4R2
Canada