Style-based distance features for author verification/profiling

Erwan Moreau [1,2] and Carl Vogel [2]

[1] Centre for Next Generation Localisation, Integrated Language Technology Group
[2] School of Computer Science and Statistics, Trinity College Dublin, Ireland

Definitions and datasets

Author identification

- Given a set of documents written by author *A* and an unknown document, find whether the latter was also written by *A*.
  - no negative evidence
- Data:
  - 35 problems with the answer: 10 in English, 20 in Greek, 5 in Spanish
  - one task = 1 to 10 *known* documents + 1 unknown document

Author profiling

- Classify a set of documents by age and gender of their author.
  - 236,000 authors in English, 75,900 authors in Spanish
  - 3 age groups $\times$ 2 genders = 6 categories

category = all the documents either in the same profile group or written by the same author.

Approach

Approach

- Inspired by the technique of **unmasking** [Koppel et al. 2007]:
  - avoid overestimating the most prominent features to capture the relevant ones,
  - but before the supervised learning stage here (as opposed to unmasking)
  - Capture the **stable** features in the category.
  - Features are **distance values** between a document and a given category.
- The frequency of frequent *n*-grams tend to follow a **normal distribution** accross documents which belong to the same category.

$\rightarrow$ use of statistics such as standard deviation, quantiles, etc.

- compare the distribution of (specific) *n*-grams in the document to the category.

Features

n-grams specific to an author/category

- **14 *n*-grams patterns** based on tokens and POS tags (15 for profiling)
  - sequential *n*-grams and skip-grams, e.g. `<token> ? <token>`
  - combinations of tokens and POS tags, e.g. `<token> <POS> <POS>`
- Various parameters to control:
  - **Representativeness.** minimum frequency of the *n*-gram in the category
  - **Consistency.** statistics on the variations of the frequency accross documents
  - **Specificity.** how the distribution of the *n*-gram differs from other categories
- Select a set *n*-grams whose frequency distributions represent the category.

Comparing a document to a category

- Classical distance measures: Euclidean, Cosine, χ^2.
- Distance measures which assume that the distribution is normal:
  - difference between this frequency and the mean in the category
  - probability to belong to this distribution with the Cumulative Distribution Function
  - ranges between quantiles
  - final value as either arithmetic, geometric or harmonic mean

Features parameters

Performance of individual features: *n*-grams patterns

- small amount of data

$\rightarrow$ POS *n*-grams/combinations better in average

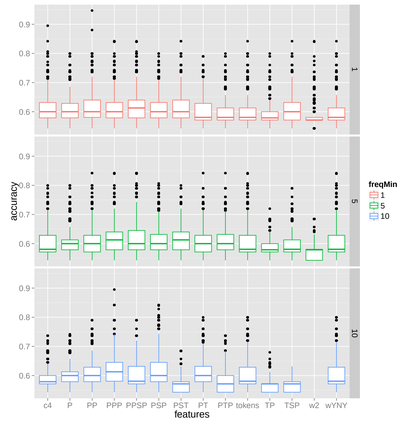
- high minimum frequency

$\rightarrow$ more stable values,

- but not enough *n*-grams

$\Rightarrow$ performance less regular

n-grams patterns: T = Token, P = POS tag, S= Skip

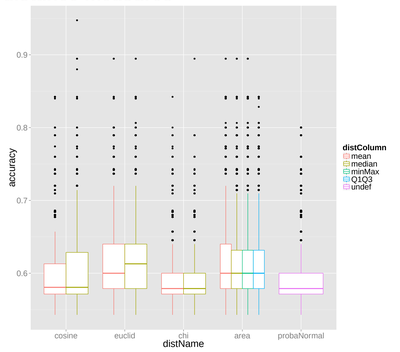


Performance of individual features: distance measures

- Euclidean distance is the best measure in average
- Measures based on CDF/PDF do not perform well
  - not enough points?

`Area`: normalized sum of the differences between the tested mean/median and one of the statistics mean/median/interquartile in the category.

`probaNormal`: Cumulative Distribution Frequency or Probability Distribution Frequency.



Experimental design

Features selection and training process

- **Distance configuration** = combination of parameters (selection + distance)
- Distance configuration applied to the task $\rightarrow$ feature value

Selection of individual distance configurations:

- Compute the features for all distance configurations
- **Incremental semi-manual selection** based on individual features
  - optimal threshold for the binary classification
  - manual analysis to assess the contribution of the parameters

$\rightarrow$ 17 best distance configurations

Final models tested by selecting **randomly**:

- A subset of *n*-grams patterns
- A distance configuration among the 17 best
- A classification algorithm with its parameters
  - SVM, Logistic Regression, Decision trees, Naive Bayes
- Evaluation using cross-validation

$\rightarrow$ For each language, select the best global configuration among $\approx$ 6000.

About negative evidence

Is negative evidence useful?

- Idea: select *n*-grams which show an unusal distribution w.r.t other categories
- Various measures tested:
  - Bhattacharrya distance (normal and general version)
  - area of the difference
  - area of the intersection
- Inconclusive results:
  - author identification: positive results, but no improvement
  - profiling: impossible to assess (bug)

Results for author identification

Language	F1-score	Best F1-score	Rank
English	0.767	0.800	3rd (tie with 1)
Greek	0.433	0.833	16th
Spanish	0.600	0.840	10th (tie with 4)
Global	0.600	0.753	11th (tie with 1)

19 participating teams.

Potential issues, future work

- Greek: possible bug with the POS tagger
- The approach might be sensitive to overfitting
  - especially when trained on a small dataset
- Semi-manual features selection process not optimal
  - predefined parameters
  - evaluation based on individual features
- Possible overlap in the information used in the two stages.

Results for author profiling

Language	Accuracy	Best accuracy	Rank
English	0.239	0.389	19
Spanish	0.254	0.421	14

21 participating teams.

Potential issues, future work

- Bug (or design flaw?) in the features selection process
- Implicit assumption: a category of authors can be analyzed as a single author
  - wrong approach is it does not hold
- Technical problems:
  - unefficient prototype $\rightarrow$ large parts of the training data ignored
  - very noisy data, basic cleaning step

Acknowledgments

This research is supported by Science Foundation Ireland (Grant 12/CE/I2267) as part of the Centre for Next Generation Localisation (www.cngl.ie) funding at Trinity College, University of Dublin.