Author Verification: Exploring a Large set of Parameters using a Genetic Algorithm

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

[1] Centre for Next Generation Localisation (CNGL)
[2] School of Computer Science and Statistics, Trinity College Dublin, Ireland

Task

Given a set of documents written by author A and an unknown document, find whether the latter was written by A.

- Output: probability in [0, 1]
- Evaluation: product of
  - Area under the ROC curve (**AUC**),
  - **c@1** (accuracy with *"don't know"* answer)

Fine-grained strategy

- Find an optimal **configuration**:
  - set of parameter/value pairs
  - methods, features, thresholds...
- Regression model based on config
  - SVM, decision trees (variants)
  - optional: confidence estimation

- **Genetic learning**
  - vast space: about 10^{19} configurations
  - maximize performance
  - risk of overfitting

- Uses a **reference corpus**
  - assuming variability among authors
  - using all documents in the dataset

Robust strategy

- A **simple distance** measure
- Words tetragrams only
- Divergence based on Jaccard sim.:

$$J_1 = \frac{(p+q)}{(p+q+r)} \qquad J_2 = \frac{(p+r)}{(p+q+r)}$$

with p words in both X and Y, q words in X but not in Y, and r words in Y but not in X

Observation types

- *n*-grams
  - tokens, characters, POS tags
  - Combinations with skip-grams
    - e.g. "`<token> __ <POS tag>`"
- stop-words *n*-grams
  - *n*-grams, only most frequent words
    - e.g. "`the __ __ is __`"
- Token length classes
  - e.g. 2-3, 3-4, 5-6, 6-7, 8-9, 10+
- Token-Type Ratio

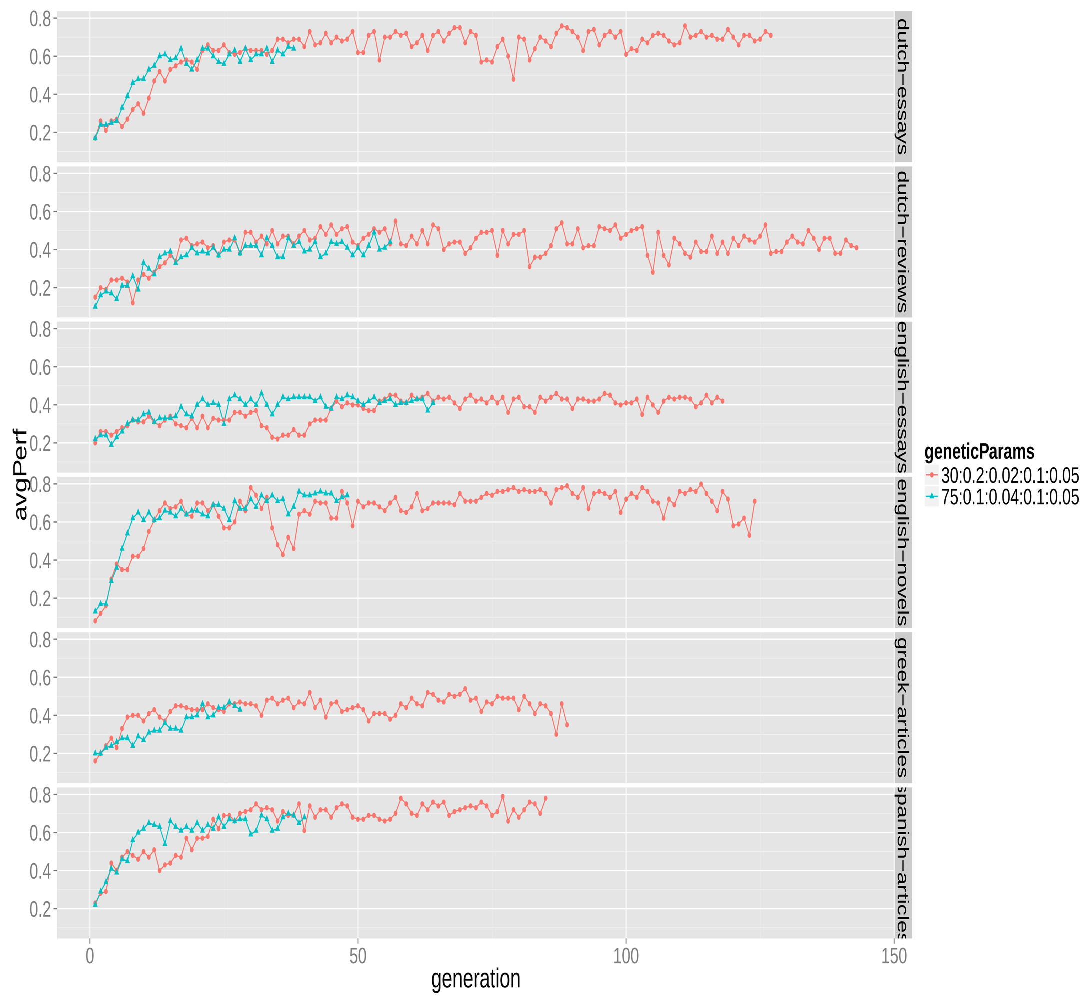
- Thresholds: min. frequency in a document, min. proportion of documents which contain the observation (known docs, ref corpus)

Features

- **Consistency**
  - how constant in known documents?
  - at least two known documents
  - std. dev., min-max range
- **Divergence**
  - how specific to the author?
  - against a reference corpus
  - mean/median diff., Bhattacharrya
- **Confidence**
  - how reliable?
  - uses consistency and divergence
- **Distance**
  - compare known vs. unknown doc
  - Cosine, Jaccard, normal distrib

Genetic algorithm

- Basic genetic learning
- Selecting configs as "breeders":
  - rank the configs by their performance
  - **better perf = higher probability**
  - pick any two breeders as parents
  - crossover, mutation
  - variants: **elitism, random**



Avg. perf. by generation, main learning stage. Parameters: population, breeders prop.; mutation probability; elitism prop.; random prop.

- **Quick convergence** in every case
- Small population sufficient
- More stable with larger population
- **14 000 to 28 000 configs evaluated**
  - main training: 3-fold cross-validation
  - final stage (best subset): 20-fold CV

Selected by the genetic algorithm

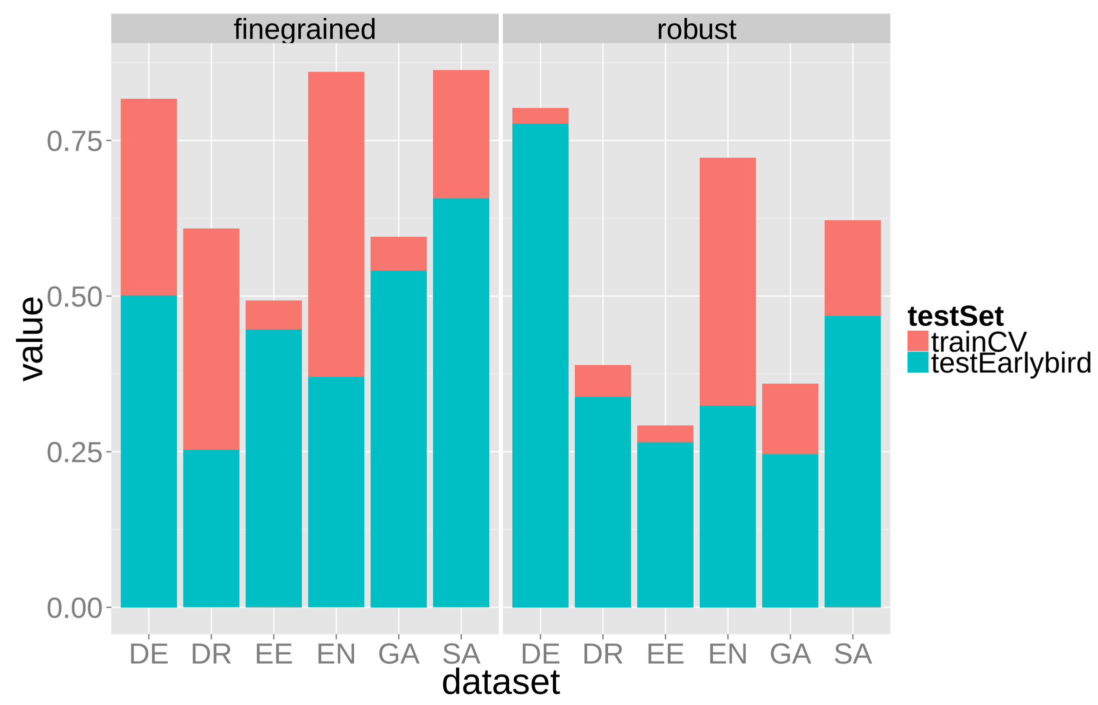
- **Observation types**
  - Few features selected (from 3 to 11)
  - POS *n*-grams, words *n*-grams
  - word length, TTR
  - stop-words *n*-grams

- **Methods**
  - Bhattacharrya (divergence measure)
  - Consistency unused in most cases
  - Simple distance metrics (e.g. cosine)
  - Decision tree regression
  - Confidence estimation unused

Selection of the final models

- Evaluation on the earlybird test set

Hypothesis: robust strategy better if only one known document?



Dataset	Known docs/case		Strategy	Perf. training	Perf. Earlybird	Perf. drop	Diff. average
Dutch essays	mean	1.79	robust	0.802	0.777	-0.025	+0.103
	median	1	fine-g.	0.817	0.501	-0.316	-0.071
Dutch reviews	mean	1.02	robust	0.389	0.338	-0.051	+0.077
	median	1	fine-g.	0.608	0.253	-0.355	-0.111
English essays	mean	2.64	robust	0.292	0.265	-0.027	+0.101
	median	3	fine-g.	0.493	0.446	-0.047	+0.198
English novels	mean	1.00	robust	0.722	0.324	-0.398	-0.270
	median	1	fine-g.	0.860	0.370	-0.490	-0.245
Greek articles	mean	2.85	robust	0.359	0.246	-0.113	+0.015
	median	3	fine-g.	0.595	0.541	-0.054	+0.191
Spanish articles	mean	5.00	robust	0.622	0.468	-0.154	-0.026
	median	5	fine-g.	0.863	0.657	-0.206	+0.039
Correlation between Diff. average and mean known docs by case						robust	0.77
						fine-g.	0.03

Results

Dataset	Training set CV		Earlybird test set			Final test set			
	robust	fine-grained	robust	fine-grained	mixed	robust	fine-grained	final	rank
Dutch essays	0.802	0.817	0.777	0.501	0.777	0.755	0.563	0.777	4
Dutch reviews	0.389	0.608	0.338	0.253	0.338	0.375	0.350	0.375	3
English essays	0.292	0.493	0.265	0.446	0.446	0.325	0.372	0.372	3
English novels	0.722	0.860	0.324	0.370	0.324	0.313	**0.352**	**0.313**	8
Greek articles	0.359	0.595	0.246	0.541	0.541	0.436	0.565	0.565	3
Spanish articles	0.622	0.863	0.468	0.657	0.657	0.335	0.634	0.634	2
Macro-average	0.531	0.706	**0.403**	**0.461**	**0.514**	**0.423**	**0.473**	**0.502**	**3**
Micro-average								**0.451**	**4**

- Hypothesis does not hold
- Selecting strategy by dataset better

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.

Computational Linguistics Group, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland   Mail: `[moreaue|jayapala|vogel]@cs.tcd.ie`   WWW: `http://www.cngl.ie`