universität
innsbruck

Dynamic Parameter Search for Cross-Domain Authorship Attribution

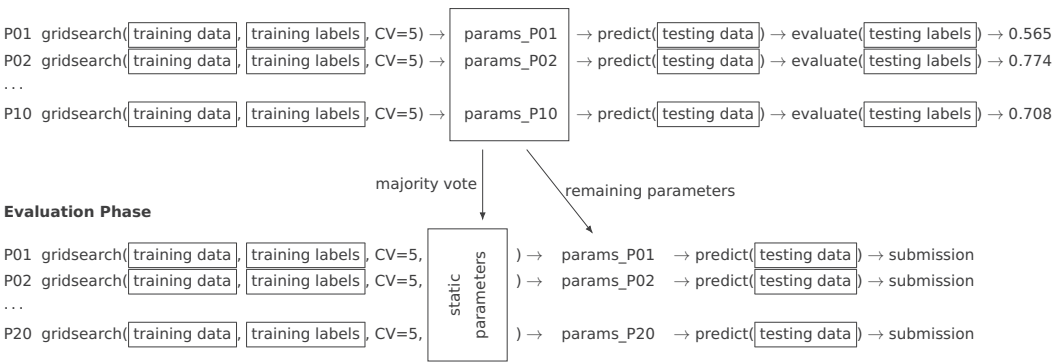
PAN @ CLEF 2018

Benjamin Murauer, Michael Tschuggnall, Günther Specht

Authorship Attribution Task

Problem	Language	Training docs	Testing docs	Avg. words/doc	Authors
p01	EN	140	105	777	20
p02	EN	35	21	782	5
p03	FR	140	49	774	20
p04	FR	35	21	782	5
p05	IT	140	80	787	20
p06	IT	35	46	807	5
p07	PL	140	103	807	20
p08	PL	35	15	788	5
p09	ES	140	117	829	20
p10	ES	35	64	851	5

Workflow


Training Phase
P01 gridsearch(training data , training labels , CV=5) →
P02 gridsearch(training data , training labels , CV=5) →
. . .
P10 gridsearch(training data , training labels , CV=5) →
params_P01
params_P02
params_P10
→ predict(testing data) → evaluate(testing labels) → 0.565
→ predict(testing data) → evaluate(testing labels) → 0.774
→ predict(testing data) → evaluate(testing labels) → 0.708
majority vote
remaining parameters
Evaluation Phase
P01 gridsearch(training data , training labels , CV=5,
P02 gridsearch(training data , training labels , CV=5,
. . .
P20 gridsearch(training data , training labels , CV=5,
static parameters
) → params_P01 → predict(testing data) → submission
) → params_P02 → predict(testing data) → submission
) → params_P20 → predict(testing data) → submission

Static Optimizations

These parameters were fixed after a preliminary, offline training run

- lowercase
- idf normalization
- keep accents
- no additional normalization (L1/L2)
- SVM C = 0.01 (tested values: 0.01, 0.1, 1, 10, 100)

Dynamic Optimizations

For these parameters, no best global option was found in the preliminary run

- minimal document frequency [0, 5]
- n-gram size [3, 4 ,5]

Results

Problem	Language	F1-score	Problem	Language	F1-score
p01	EN	0.73	p11	IT	0.841
p02	EN	0.689	p12	IT	0.534
p03	EN	0.8	p13	PL	0.473
p04	EN	0.83	p14	PL	0.527
p05	FR	0.55	p15	PL	0.369
p06	FR	0.608	p16	PL	0.432
p07	FR	0.609	p17	ES	0.631
p08	FR	0.662	p18	ES	0.771
p09	IT	0.659	p19	ES	0.783
p10	IT	0.616	p20	ES	0.75
		Average	0.643		

Final Rank

Contestant	Mean F1-score
custodio18	0.685
murauer18	**0.643**
halvani18	0.629
...	
baseline	0.584

Some Failed Attempts

- Word and document embeddings and CNN
  $\rightarrow$ failed to find pre-trained data for all languages
  $\rightarrow$ failed to create custom embedding
- Word-based n-grams
  $\rightarrow$ performance drop
- POS-based featuers
  $\rightarrow$ promising for English, failed to find good parsers for all languages
  $\rightarrow$ (too) high memory requirements