UniNE at CLEF 2015: Author Identification

Mirco Kocher, Jacques Savoy

University of Neuchâtel, Switzerland

Swiss National Science Foundation

Task

Given a small set of "known" documents (no more than seven, possibly as few as one) written by a single person, is the new "unknown" document also written by that author?

- Output: probability of authorship
- Evaluation: final score is product of
  - Area under the ROC curve
  - c@1 (accuracy with "*don't know*" answer)
- Different datasets:
  - Dutch and Spanish – cross-genre corpora
  - English and Greek – cross-topic corpora
- *TIRA* platform for evaluation
  - Encapsulated system with restricted data access
  - Fair comparison of the time needed to produce an answer
- Total 18 different participants

Strategy

Term	unknown	known	cand1	cand2	cand3
the	0.09	0.09	0.09	0.10	0.06
of	0.06	0.06	0.06	0.06	0.04
to	0.03	0.04	0.00	0.01	0.05
and	0.02	0.03	0.04	0.02	0.00
…					

200

m = 3, randomly selected

- L_1-norm as distance measure:
  - unknown – known: 0.02 -> candidate Δ_0
  - unknown – cand1: 0.05
  - unknown – cand2: 0.03 -> first impostor Δ_{m1}
  - unknown – cand3: 0.09
- Repeat r times and compute arithmetic mean of $\Delta_{m1}, \ldots, \Delta_{mr}$
  - r = 5, gives the final impostor difference $\Delta_{\bar{m}}$
- Decide according to ratio $\Delta_0 / \Delta_{\bar{m}}$ with a threshold of 2.5%
  - if $\Delta_0 / \Delta_{\bar{m}} < 0.975$ then "*same author*"
  - if $\Delta_0 / \Delta_{\bar{m}} > 1.025$ then "*different author*"
  - otherwise "*don't know*"
- Assign probabilities:
  - Fit all "*different author*" answers to range [0.0, 0.5)
  - 0.5 for "*don't know*" answers
  - Fit all "*same author*" answers to range (0.5, 1.0]

Results

Language	Final	AUC	c@1	Runtime	Rank
Dutch	0.2175	0.4495	0.4840	00:00:07	14
English	0.5082	0.7375	0.6890	00:00:24	4
Greek	0.6310	0.8216	0.7680	00:00:11	3
Spanish	0.3665	0.6498	0.5640	00:00:22	10

- Compared with all 18 participants:
  - Good scores for Greek and English (both cross-topic)
  - Low scores with Dutch and Spanish (both cross-genre)
  - Fast runtime (~1 minute), median execution time of other systems is almost one hour (excluding their training time)

Evaluation

- Probabilistic approach:
  - Possible variation around the reported performance is small
  - *E.g.* English mean c@1 is 0.5776 with σ=0.0237
  - Standard deviation based on 200 restarts with random impostor selection
- Limit of 2.5%:
  - Lower: forces the system to decide without enough evidence
  - Higher: encourages the classifier not to take a decision
- Vary *m* (number of impostors) and *r* (number of iterations) from 1 to 7
  - No significant difference when using the best combination
- Explanation of proposed assignment (*e.g.* in English corpus):
  - Usually, the relative frequency differences with very frequent words such as *when*, *is*, *in*, *that*, *to*, or *it* can explain the decision

Conclusion

- Multiple languages with the genre and topic differ significantly
- Macro-averaging ranks us 8th out of 18 participants
- Unsupervised approach
- No language dependent adjustments or parameter training
- Simple technique can solve the verification problem
- Most frequent terms tend to select most discriminative features
- Reduced set of comprehensible features could clearly explain the decision taken
- Genre variation is difficult to handle without fixing some parameters
- Knowing the degree of belief is an important aspect

Acknowledgments

This research was supported, in part, by the NSF under Grant #200021_149665/1.

Mirco Kocher, Jacques Savoy – University of Neuchâtel – Rue Emile-Argand 11 – 2000 Neuchâtel, Switzerland – {Mirco.Kocher, Jacques.Savoy}@unine.ch

http://style.unine.ch/