Author Verification
Using Common N-Gram Profiles of Text Documents

Magdalena Jankowska, Evangelos Milios, and Vlado Kešelj

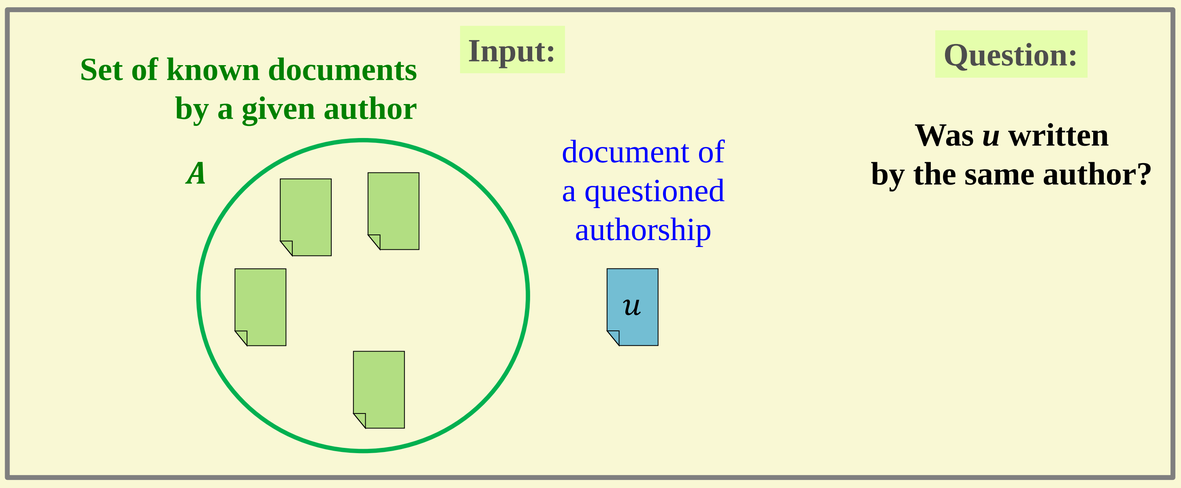
Faculty of Computer Science, Dalhousie University, Halifax, Canada
{jankowsk, eem, vlado} @cs.dal.ca

25th International Conference on Computational Linguistics, COLING 2014

Our approach to the Author Verification problem

Author Verification problem
Given a few samplings (possibly just one) of a person's writing, was a questionable document written by the same person?

Applications
- Forensics
- Security
- Literary research



Our approach

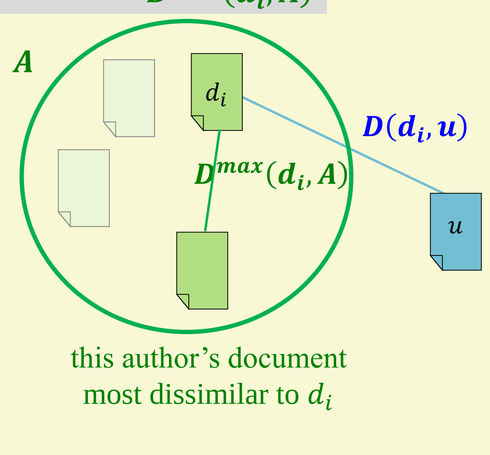
A **proximity based one-class classification method** exploiting pairs of most dissimilar documents of the given authorship.

For the document dissimilarity we applied **CNG (Common N-Gram) dissimilarity** (Kešelj et al. 2003). A document is represented by a *profile:* a list of its most common n-grams (or characters or words) with their normalized frequencies.

Ensembles combine classifiers that differ between each other with respect to at least one of the three document representation parameters: type of the tokens for n-grams (word or character), size of n-grams, length of a profile.

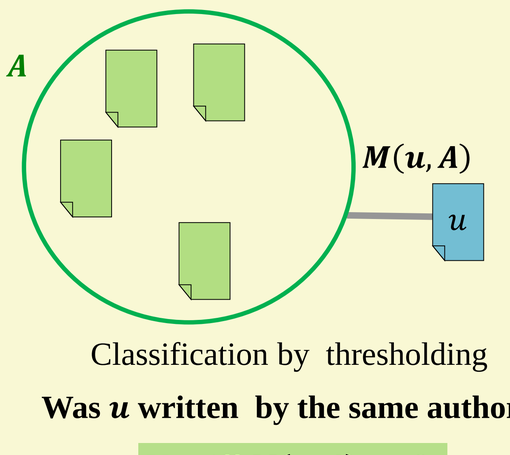
dissimilarity ratio of a "known" document d_i
measure of how much more/less dissimilar is the questioned document than the most dissimilar document by this author.

$$r(d_i, u, A) = \frac{D(d_i, u)}{D^{max}(d_i, A)}$$



this author's document most dissimilar to d_i

Proximity between the questioned document and the set of known documents
$M(u, A)$ - average of dissimilarity ratios $r(d_i, u, A)$ over all "known" documents d_i



Classification by thresholding

Was u written by the same author?

Yes, iff $M(u, A) < \theta$

Evaluation on the dataset of PAN'13 Author Identification competition

Ensembles evaluation (after the competition)
- competitive (as compared to the best competition results) results on the entire set, and the English and Spanish subsets
- especially good results by ensembles combining character-based and word-based classifiers

Our **competition submission** of a single classifier:
ranking 5th (joint) of 18 according to accuracy,
ranking 1st of 10 according to AUC.

Selected experimental results	Entire set		English subset		Spanish subset		Greek subset	
	accuracy	AUC	accuracy	AUC	accuracy	AUC	accuracy	AUC
Our single classifier with parameters tuned on training data								
competition submission	0.682	0.793	0.733	0.839	0.720	0.859	0.600	0.711
Our ensembles: weighted voting, all classifiers in the considered parameter space								
character based	0.729	0.764	**0.833**	0.830	0.800	0.859	0.567	0.582
character and word based	0.741	0.780	0.800	0.842	**0.840**	0.853	0.600	0.622
Our ensemble: weighted voting, classifiers selected based on performance on training data								
character and word based	**0.788**	**0.805**	0.800	**0.857**	**0.840**	0.853	0.733	0.687
Methods by other PAN'13 participants (different methods in different columns)								
best results over other participants	0.753	0.735	0.800	0.837	**0.840**	**0.926**	**0.833**	**0.824**

Conclusions and Future Work

Conclusions
- proximity based one-class classification method with promising results for authorship verification
- requires at least two samples of writing by a given author
- combining word n-gram based and character n-gram based classifiers yields best results

Future research directions
- better adaptation of the method for the case of a single document of the known authorship
- analysis of the role of word n-grams and character n-grams depending on the genre of the texts, and on the topical similarity between the documents

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