INSTITUTO NACIONAL DE ASTROFÍSICA, ÓPTICA Y ELECTRÓNICA

EPSMS and the Document Occurrence Representation for Authorship Identification

Hugo Jair Escalante[1], Manuel Montes[2]
1 Universidad Autónoma de Nuevo León,
2 Instituto Nacional de Astrofísica, Óptica y Electrónica
hugo.jair,@gmail.com, mmontesg@inaoep.mx

For the **authorship attribution** task we performed experiments with a document occurrence representation using a standard classification-based approach. Results obtained with this approach were mixed: in the small data sets distributional representations resulted very helpful, although in the large data sets a simple bag-of-words approach outperformed the document occurrence approach. For the **authorship verification** task we adopted a classification-based approach and proposed a modification to Ensemble Particle Swarm Model Selection (EPSMS) for selecting classification models for each task. This approach obtained acceptable performance in two out of the three data sets.

Authorship attribution

A neural network classifier was used under the one-vs-all formulation. Documents were represented under the document occurrence representation (DOR).

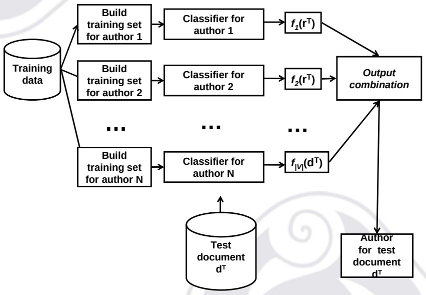

Fig.1. OVA approach to authorship attribution

The document occurrence representation

DOR is a distributional term representation in which a document is represented by a distribution of occurrences over other documents in the same corpus. Intuitively, a document is represented by its context. Each term in the vocabulary is first represented as a distribution of occurrences over documents.

$$\mathbf{w}^{dor}(t_j,\mathbf{d}_k) = df(t_j,\mathbf{d}_k) \times \log(\frac{|V|}{N_k})$$

Next, each document is then represented by a combination of the representations of terms that occur in the document.

$$df(t_j,\mathbf{d}_k) = \begin{cases} 1+\log(\#(t_j,\mathbf{d}_k)) & if\ \#(t_j,\mathbf{d}_k) > 0 \\ 0 & otherwise \end{cases}$$

Experimental results at PAN’11

Dataset	MA-P	MA-R	MA-F1	MI-P	MI-R	MI-F1	Sum-ranks	Overall rank
Small	0.676	0.381	0.387	0.709	0.709	0.709	19	3rd out of 17
Small+	0.65	0.201	0.193	0.578	0.573	0.575	16	2nd out of 13
Large	0.608	0.294	0.303	0.508	0.508	0.508	48	8th out of 18
Large+	0.53	0.203	0.191	0.446	0.446	0.446	29	5th out of 13

Table 1. Official results obtained with the proposed approach.

Results were very competitive, the entries were above the average performance among other participants. Results were particularly positive in the Small data sets and for the “+” version.

Does DOR really help?

Dataset	Accuracy		Macro-F1		Micro-F1		Sum-ranks	
	DOR	BOW	DOR	BOW	DOR	BOW	DOR	BOW
Small	**70.91**	67.88	0.387	**0.418**	**0.709**	0.678	3	4
Small+	**57.25**	55.20	0.709	**0.552**	0.193	-	2	2
Large	50.76	**62.53**	0.303	**0.463**	0.507	**0.625**	8	3
Large+	44.56	**53.24**	0.446	**0.532**	0.191	-	5	2

Table 2. Comparison of the DOR and bag-of-words (BOW) representations.

We compared the performance of the DOR representation with a straight BOW formulation. DOR outperformed BOW for the Small data sets (accuracy). Although, BOW outperformed DOR in the Large datasets.

Authorship verification

We faced the verification problem as a classification task (docs. written by the author vs. docs. written by any other author). Documents were represented with the BOW formulation and EPSMS was used to select the classifier .

Ensemble Particle Swarm Model Selection

EPSMS is a method for the automatic selection of binary classification models. In a nutshell, EPSMS searches for the best ensemble method that can be generated by using the methods available in a machine learning toolbox.

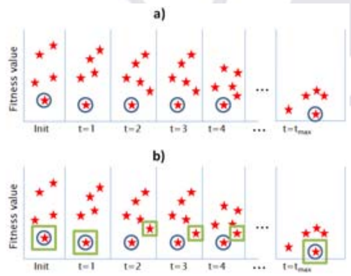

Fig.2. Illustration of PSMS (a) and EPSMS(b) approaches. In PSMS particle swarm optimization is used to explore the space of full models represented as stars (a). A full model is that composed of methods for preprocessing feature selection and classification. EPSMS selects a subset of models evaluated during the search and uses them to build an ensemble classifier (b).

In order to obtain stable predictions, the results obtained with five runs of EPSMS were combined to obtain the final prediction of the model.

Experimental results at PAN’11

Dataset	MA-P	MA-R	MA-F1	MI-P	MI-R
Verify-1	0.1	0.333	0.154	17	6th out of 10
Verify-2	0.4	0.8	0.533	11	1st out of 10
Verify-3	0	0	0	30	9th out of 10

Table 3. Official verification results obtained with the proposed approach.

The results obtained with the proposed formulation are interesting and give evidence that the classification approach to AV can be very effective. We believe the proposed method has potential for this and other tasks, although we would like to conduct an extensive evaluation of it in order to detect what factors influence the performance of the proposed technique.

Conclusions

❑ For AA we found that the use of DOR was partially beneficial in a classification-based approach.

❑We confirmed that the BOW is a strong baseline.

❑In general, the classification-based approach to AA and AV is very competitive nowadays.

❑ In AV we found EPSMS is competitive although it still can further improved. We are studying was of tailoring EPSMS for the AA and AV tasks.

References

[1] A. Lavelli, and F. Sebastiani, and R. Zanoli. Distributional Term Representations: An Experimental Comparison. Proc. of the International Conference of Information and Knowledge Management, pp. 615—624, ACM, 2005.

[2] H. J. Escalante, M. Montes, E. Sucar. Ensemble Ensemble Particle Swarm Model Selection. Proc. of the World Congress on Computational Intelligence, pp. 1814—1821, IEEE, 2010