Representation learning of writing style

Julien Hay, Bich-Liên Doan, Fabrice Popineau and Ouassim Ait Elhara

We introduce a new method of representation learning that aims to embed documents in a **stylometric space**. Previous studies in the field of authorship analysis focused on **feature engineering** techniques in order to represent document styles and to enhance model performance in specific tasks. Instead, we directly embed documents in a stylometric space by relying on a **reference set of authors** and **the intra-author consistency property** which is one of two components in our definition of writing style. The method we propose allows for the clustering of documents based on stylistic clues reflecting the authorship of documents. For the empirical validation of the method, we train a **deep neural network model** to predict authors of a large reference dataset consisting of news and blog articles. Albeit the learning process is supervised, it does not require a dedicated labeling of the data but it relies only on the metadata of the articles which are available in huge amounts. We evaluate the model on multiple datasets, on both the **authorship clustering** and the **authorship attribution** tasks.

Method

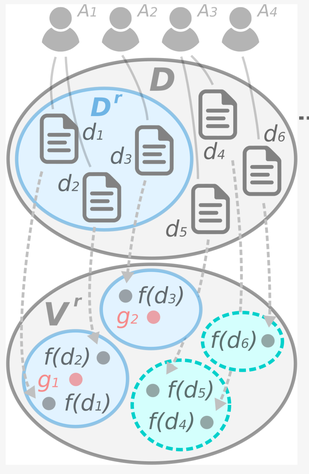
How to learn style representation and avoid feature engineering?

Learn a function *f* able to embed documents by authorship on the basis of a large reference corpus following the *style-generalization* assumption. By using, for instance, a DNN with attention layers, we seek to focus on latent structures satisfying:

- **The intra-author consistency**: the property of being consistent in documents belonging to the same author [2, 3].
- **The semantic undistinguishness**: the property of carrying very little information on what makes the document semantically (e.g. topics, named entities) distinguishable in the corpus [4].

The *style-generalization* assum.

Given *f* able to accurately cluster documents by reference authors in D^r, the ***style-generalization* assumption** states that two representations defined by *f* of two **unseen documents** (d4 and d5) are more likely to be similar if they belong to the same author than if they do not. Intuitively, we assume that any unseen document belonging to an unknown author is similar, in terms of style, to documents belonging to a **subset of known authors** and that another document of the same unknown author is likely to be similar, in terms of style, to documents belonging to this **same subset** of known authors.



Assessing the intra-author consistency

Using a standard internal clustering index, the **Davies–Bouldin index** [1]:

$$DavB = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\frac{s_i + s_j}{d_{ij}}$$

And a new metric called **SimRank**:

$$\mathrm{SimRank}(REL) = \frac{\sum_{p=1}^{|REL|}\mathrm{nDCG}'(REL_p)}{|REL|}$$

which is normalised, stable for different samples having the same cluster configuration and is able to handle classes having multiple clusters. This corresponds more to our case when the same author can adopt multiple writing styles.

Id	Clusters	SimRank (±2σ)	DavB (±2σ)
1		1.0 (±0)	0.25 (±0.01)
2		0.69 (±0.01)	170 (±393)
3		0.66 (±0)	1.27 (±0.05)
4		0.53 (±0)	38 (±109)

Assessing the semantic undistinguishness

For models with attention layers, we look at attention weights and words TFIDF weights by using the **TFIDF Focus** measure:

$$\mathrm{TFIDFFocus}(A,T) = \frac{\sum_{i=1}^{d}\sum_{j=1}^{w} A_{ij}.T_{ij}}{d}$$

Experiments

Reference corpus and test sets

We split a large corpus of news and blog articles into a ***R-set*** (the reference corpus) and multiple ***U-sets*** having unseen documents (from the point of view of *f*). Each *U-set* is composed of 2500 documents from 50 writers of either: *The Washington Post*, *Breitbart*, *Business Insider*, *CNN*, *The Guardian*, *The New York Times*, etc. We used **22 *U-sets*** in these experiments.

Learning the *f* function

In order to learn *f*, we choosed to train a **BERT model** [5] (pre-tained using unsupervised learning) with attention layers to predict authors in the *R-set*, and take **intermediate weights** in the DNN as representations of unseen documents from *U-sets*.

We compared our model to standard document representation models such as **topic models** (e.g. *LDA*), bag-of-words models (e.g. *TFIDF*) and **representation learning models** (e.g. *Doc2Vec*).

Model	SimRk	DavB	Acc
Random	0.185	14.81	/
TFIDF	0.455	**4.683**	0.514
LDA	0.309	8.353	0.163
Stylo	0.276	65.71	0.098
Doc2Vec	0.430	6.194	0.472
USent	0.416	5.328	0.499
InferSent	0.374	5.625	0.594
BERT	0.378	5.469	0.536
SNA	0.463	4.785	0.552
DBert	0.339	7.058	0.522
DBert-ft	**0.474**	4.777	**0.597**

Results

- Our model (*DBert-ft*) performs well on the **clustering experiment** according to SimRank scores and DavB scores (close to the standard BOW model, *TFIDF*).
- On an **external task**, the authorship attribution task (classification with the accuracy metric), our model also performs well (close to *InferSent*). We used document representation from each model and a SVM to identifiy authors. Average performances on both tasks show the benefits of our method.
- The **semantic undistinguishness adequacy** of our model is assessed using the *TFIDF Focus* measure. We quantitatively showed that our method allows, for the trained models, to **focus their attention on function words** having lower TFIDF weights (10% less on average). Thus, we may conclude that our model is more able to capture stylometric features related to specific words exposing the semantic undistinguishness property than other models trained on smaller *U-sets*.

U-set type	*SNA* trained on	*U-set 1*	*U-set 2*	*U-set 3*	*U-set 4*	*U-set 5*	*Mean*
News	***Target U-sets***	0.642	0.650	0.606	0.601	0.661	0.632
	Other U-set	0.611	0.591	0.576	0.559	0.623	0.592
	R-set	**0.497**	**0.479**	**0.477**	**0.459**	**0.507**	**0.483**
Blog	***Target U-sets***	0.668	0.722	0.734	0.645	0.702	0.694
	Other U-set	0.637	0.670	0.670	0.606	0.648	0.646
	R-set	**0.547**	**0.579**	**0.575**	**0.526**	**0.560**	**0.557**

[1] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure", in IEEE Transactions on Pattern Analysis and Machine Intelligence (1979)
[2] David I. Holmes, "The Evolution of Stylometry in Humanities Scholarship", in Literary and Linguistic Computing (1998)
[3] Jussi Karlgren, "The wheres and whyfores for studying text genre computationally", in Workshop on Style and Meaning in Languange, Art, Music and Design.
[4] Efstathios Stamatatos, "Masking topic-related information to enhance authorship attribution", in Journal of the Association for Information Science and Technology (2018)
[5] Sanh et al., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", in NeurIPS EMC2 Workshop (2019)